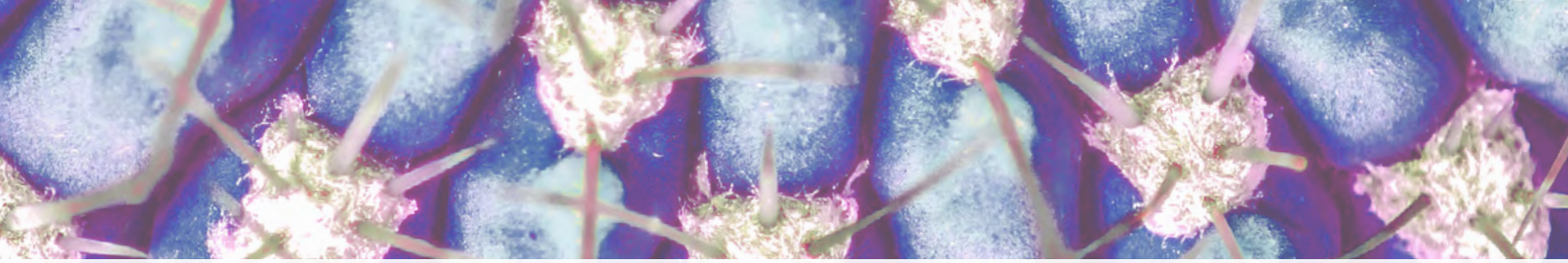

ENAR 2016 Abstracts & Poster Presentations

1. POSTERS: LATENT VARIABLES AND MIXTURE MODELS

1a. INVITED POSTER: THE LZIP: A BAYESIAN LATENT FACTOR MODEL FOR CORRELATED ZERO-INFLATED COUNTS

Brian Neelon*, Medical University of South Carolina

Dongjun Chung, Medical University of South Carolina

Motivated by a study of molecular differences among cancer patients, we develop a Bayesian latent factor zero-inflated Poisson (LZIP) model for the analysis of correlated zero-inflated counts. The responses are modeled as independent zero-inflated Poisson distributions conditional on a set of subject-specific latent factors. For each outcome, we express the LZIP model as a function of two discrete random variables: the first captures the propensity to be in an underlying "at-risk" state, while the second represents the count response conditional on being at risk. The latent factors and loadings are assigned conditionally conjugate gamma priors that accommodate overdispersion and dependence among the outcomes. For posterior computation, we propose an efficient data augmentation algorithm that relies primarily on easily sampled Gibbs steps. We conduct extensive simulations to investigate both the inferential properties of the model and the computational capabilities of the proposed MCMC algorithm. We apply the method to an analysis of breast cancer genomics data from the Cancer Genome Atlas.

email: *neelon@musc.edu*

1b. UNDERSTANDING GAUSSIAN PROCESS FITS USING AN APPROXIMATE FORM OF THE RESTRICTED LIKELIHOOD

Maitreyee Bose*, University of Minnesota

James S. Hodges, University of Minnesota

Sudipto Banerjee, University of California, Los Angeles

A Gaussian process (GP) is often used as the random effect in a linear mixed model, with its unknowns estimated by maximizing the log restricted likelihood or doing a Bayesian analysis. 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 the GP parameters are fit to data by deriving a simple, interpretable, and fast-computing form of the restricted likelihood. We have applied the spectral approximation to the intercept-only GP; the log restricted likelihood arising from this approximation has a simple form, which 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. We then introduce covariates into the model, and present analysis of California asthma hospitalization data using our GLM representation as the basis for a diagnostic tool to identify missing covariates. Throughout, only finite-sample methods are used.

email: *bosex020@umn.edu*

1c. A JOINT DISTRIBUTION FOR A TIME-TO-EVENT OUTCOME AND RECURRENT EVENTS

Luojun Wang*, Penn State University

Vernon M Chinchilli, Penn State University

Recurrent events are commonly encountered in biomedical research studies and clinical trials. When recurrent events are correlated with a failure event, such as death, we no longer should assume independent censoring. Many reports in the literature incorporate a latent variable model to account for the correlation between the time to event T and the number of recurrent events N(t). In a fully parametric setting where full likelihood-based analyses can be applied, however, such a model may be over-parameterized and difficult to apply. Here, we propose a joint distribution for (T, N) based on conditional distributions. We illustrate the use of this joint distribution to model the recurrent events of acute kidney injury (AKI) and time to primary outcome (death) in patients with and without chronic kidney disease (CKD). In this fully parametric model, we develop the intensity ratio for the recurrent events and the hazard ratio for the failure event among different groups of patients with or without an AKI event at baseline and with or without CKD at baseline. Based on our model, we then investigate if recurrent AKI is predictive of death.

email: *vicwong@psu.edu*

1d. ANALYSIS OF PHIS DATA FOR A ZERO-TRUNCATED, 1&2 INFLATED, AND MULTI-LEVEL COUNT VARIABLE

Ji Young Kim*, The Children's Hospital of Philadelphia

Benjamin L. Laskin, The Children's Hospital of Philadelphia

Tamar Y. Springel, University Hospital

Susan L. Furth, The Children's Hospital of Philadelphia and University of Pennsylvania

Justine Shults, University of Pennsylvania

WITHDRAWN

The Pediatric Health Information System (PHIS) database contains clinical and resource utilization data from 45 children's hospitals in the US. The data includes admission, diagnosis, treatment, as well as administrative details such as discharge and length of hospitalization. Each patient in the database is tracked over time. This type of data features multi-level correlations, where longitudinal observations are nested under patients, and patients are nested under hospitals. Hospitalization length of stay is a readily examined variable in the PHIS data analyses. However, when it comes to modeling the count data such as the hospitalization length of stay (in days), the analysis may encounter interesting challenges of simultaneously dealing with the zero-truncated, 1&2 inflated and multi-level correlated structures. In this presentation, we will discuss how we can approach to these challenges with an application to the PHIS data, in the context of regression.

email: *kimj15@email.chop.edu*

1e. A LATENT VARIABLE APPROACH TO ELICIT CONTINUOUS TOXICITY SCORES AND SEVERITY WEIGHTS FOR MULTIPLE TOXICITIES IN DOSE-FINDING ONCOLOGY TRIALS

Nathaniel S. O'Connell*, Medical University of South Carolina

Elizabeth Garrett-Mayer, Medical University of South Carolina

The goal of most dose-finding oncology trials is to find the highest dose yielding an acceptable patient toxicity profile. Conventional dose-finding trials utilize binary toxicity endpoints that treat low to moderate toxicities as irrelevant, ignoring potentially harmful combinations of such toxicities. Addressing these concerns, novel methods have been introduced for composite toxicity scores accounting for multiple toxicities of varying grades, but calculation of such scores require prior specification of severity weights representing the relative toxicity burden each observed toxicity contributes to a toxicity profile. Elicitation of weights generally rely on subjective specification, and resulting scores may be confusing in clinical settings. We propose a novel method of estimating toxicity weights via a cumulative logit model, assuming a latent continuous toxicity score characterized by the set of observed toxicity types and grades, linked to ordinal outcomes corresponding to intuitive dose escalation decisions. The toxicity score elicitation method (TSEM) produces an accurate scoring system through evaluation of a balanced subset of toxicity profiles in terms of severity, and we present an adaptive weight finding algorithm to facilitate this. The TSEM bridges the gap between relating toxicity scores to clinically interpretable outcomes, and provides an objective method for determining toxicity weights and scores.

email: *oconneln@musc.edu*

1f. MULTILEVEL BINARY PRINCIPAL COMPONENT ANALYSIS

Yuting Xu*, Johns Hopkins Bloomberg School of Public Health

Chen Yue, Johns Hopkins Bloomberg School of Public Health

Vadim Zipunnikov, Johns Hopkins Bloomberg School of Public Health

Martin A. Lindquist, Johns Hopkins Bloomberg School of Public Health

Brian S. Caffo, Johns Hopkins Bloomberg School of Public Health

Principle component analysis (PCA) has been widely used in data decomposition and dimension reduction. In this work, we developed a multilevel probabilistic PCA model to analyze high-dimensional binary data with replicate measurements. We use a generalized linear model to describe the mean structure and decompose the subject-level and subject-replicates-level deviations using the multilevel probabilistic PCA framework. Our goal is to find between-subject level principal components (PCs) as well as within-subject level PCs. We proposed a nested variational EM approach for model fitting, which is computationally efficient and scalable for large data sets. The application to simulated data as well as NHANES physical activity data illustrates the efficacy of proposed method.

email: *xuyuting@jhu.edu*

1g. A SCORE TEST FOR DETECTING PUBLICATION BIAS IN MULTIVARIATE RANDOM-EFFECTS META-ANALYSIS

Chuan Hong*, University of Texas Health Science Center, Houston

Haitao Chu, University of Minnesota

Yong Chen, University of Pennsylvania Perelman School of Medicine

Publication bias occurs when the publication of research results depends not only on the quality of the research but also on the direction, magnitude, or statistical significance of the results. Publication bias may threaten the validity of systematic reviews and meta-analyses. Multivariate meta-analysis has recently received increasing attention for its potential ability to reduce bias and improve statistical efficiency by borrowing information across outcomes. However, both detecting and accounting for publication bias are more challenging in a multivariate meta-analysis setting. In this

paper, we propose a pseudolikelihood-based score test for detecting publication bias in multivariate random-effects meta-analysis. To the best of our knowledge, this is the first test for detecting publication bias in a multivariate meta-analysis setting. Two detailed case studies are given to show the limitations of univariate tests and to illustrate the advantage of the proposed test in practice. Through simulation studies, the proposed test is found to be more powerful than the existing univariate tests. In addition, by empirically evaluating 169 systematic reviews with multiple outcomes from the Cochrane Database, the proposed multivariate test is shown to identify more studies with publication bias than existing univariate tests.

email: *chuan.hong@uth.tmc.edu*

1h. MEASURING CONCURRENCY USING A JOINT MULTISTATE AND POINT PROCESS MODEL FOR RETROSPECTIVE SEXUAL HISTORY DATA

Hilary J. Aralis*, University of California, Los Angeles

Pamina M. Gorbach, University of California, Los Angeles

Ron Brookmeyer, University of California, Los Angeles

Understanding the impact of concurrency, defined as overlapping sexual partnerships, on the spread of HIV within various communities has been complicated by difficulties in measuring concurrency. Previous attempts to empirically estimate the magnitude and extent of concurrency among surveyed populations have inadequately accounted for the dependence between partnerships and used only a snapshot of the available data. We introduce a joint multistate and point process model in which states are defined as the number of ongoing partnerships an individual is engaged in at a given time. Sexual partnerships starting and ending on the same date are referred to as one-offs and modeled as discrete events. The proposed method treats each individual's continuation in and transition through various numbers of ongoing partnerships as a separate stochastic process and allows the occurrence of one-offs to impact subsequent rates of partnership formation and dissolution. Estimators for the concurrent partnership distribution and mean sojourn times are presented. We demonstrate this modeling approach using epidemiological data collected from a sample of men having sex with men and seeking HIV testing at a Los Angeles clinic. Among this sample, the estimated point prevalence of concurrency was higher among men later diagnosed HIV positive.

email: *hilary.aralis@gmail.com*

1i. EVALUATING QUALITY OF WEB PANEL SURVEY DATA VIA CLUSTERING AND LATENT CLASSES

Elizabeth Handorf*, Fox Chase Cancer Center, Temple University

Susan Darlow, National Comprehensive Cancer Network

Michael Slifker, Fox Chase Cancer Center, Temple University

Carolyn Heckman, Fox Chase Cancer Center, Temple University

Lee Ritterband, University of Virginia

Online surveys using web panels are a valuable tool for social science research, however, the perceived anonymity provided by online enrollment and study completion may lead to inaccurate responding. Particularly, in a study offering incentives, without certain protections in place, participants may enroll multiple times, or as in any survey-based research, participants may be careless or inaccurate in their answers. In this work, we illustrate a method to identify poor-quality responses motivated by an analysis of survey data which measure the effectiveness of an online skin-cancer prevention program. To identify repeat enrollees, we constructed a Euclidean distance matrix based on responses to a series of eligibility questions, and then performed hierarchical clustering with complete linkage to identify spherical clusters. We assessed 19 potential quality indicators based on participant registration information and baseline survey responses, including previously proposed and study-specific measures. These quality indicators were combined with cluster membership in a latent class model, and responses were categorized as high or low quality based on their fitted latent class membership. As expected, we find that removing low-quality responses yields larger estimates of the treatment effect at follow-up. This analysis demonstrates that including poor-quality responses may bias study results towards the null.

email: *elizabeth.handorf@fccc.edu*

1j. JOINT MODELING OF LONGITUDINAL, RECURRENT EVENTS AND FAILURE TIME DATA FOR SURVIVOR'S POPULATION

Qing Cai*, Johns Hopkins University

Mei-Cheng Wang, Johns Hopkins University

Gary Chan, University of Washington

Recurrent events together with longitudinal measurements are commonly observed in follow-up studies where the observation is terminated by censoring or a primary failure event. In this paper we developed a joint model where the correlation of longitudinal measurements, recurrent event process and time to failure event is modeled through rescaling the time index. The general idea is that the trajectories of all biology processes of subjects in the survivors' population are elongated or shortened by the rate identified from a failure time model for the failure event. The model is constructed on the basis of survivors' population which avoids making disputing assumption on recurrent events or biomarkers after the failure event (such as death). The model also possesses a specific feature that, by aligning failure events as time origins, the backward-in-time model of recurrent events and longitudinal measurements shares the same parameter values with the forward time model. The

statistical properties, simulation studies and real data examples are conducted. The proposed model and estimation inference can be generalized to analyze left-truncated data.

email: *qcai3@jhu.edu*

1k. MODEL DIAGNOSTICS AND PREDICTIVE POWER ASSESSMENT OF A TYPE OF JOINT DYNAMIC MODELS OF RECURRENT COMPETING RISKS AND A TERMINAL EVENT

Piaomu Liu*, University of South Carolina, Columbia

Liu and Peña (2015) proposed a type of joint dynamic modeling of recurrent competing risks (RCR) and a terminal event (TE) for both the no-frailty and frailty cases. Counting processes were assumed as the underlying data generation mechanism. The models considered impact of past recurrent competing event occurrences, possible interventions and covariate information. Moderate simulations showed that the proposed semiparametric estimation procedures perform reasonably well. Individual dynamic prediction of terminal event times has also been proposed. To better understand properties of the joint models, in this paper, we examine model fit and predictive power of the models. Martingale-based residuals are used in assessing model fit. Both ad-hoc procedures and a type of mean squared error (MSE) are proposed to understand the predictive power of the models.

email: *piaomuliu@gmail.com*

2. POSTERS: IMAGING AND SPATIOTEMPORAL APPLICATIONS

2a. INVITED POSTER: BIG DATA AND NEUROIMAGING: LARGE-SCALE MODELS FOR BRAIN NETWORKS

Xi Luo*, Brown University

The human brain contains billions of interconnected neurons. Intrinsically, neuroimaging data are big data and with complex structures. This poster introduces an optimization approach based on graphical models to infer large-scale brain networks with up to millions of nodes. We introduce a framework of hierarchical networks. This framework provides simultaneous dimension reduction and network edge estimation. We formulate this framework as a conditional convex optimization problem, with an innovation of separating the complex optimization criterion into many computationally tractable ones. In addition to better estimation accuracy, our approach also provides advantages in interpreting and visualizing such large networks. These advantages will be illustrated using simulations and real fMRI datasets

email: *xi.rossi.luo@gmail.com*

2b. SCALAR ON IMAGE REGRESSION WITH APPLICATION TO MULTIPLE SCLEROSIS MRI LESION DATA

Cui Guo*, University of Michigan

Timothy D. Johnson, University of Michigan

Multiple sclerosis (MS) is an autoimmune disease that attacks the central nervous system. In particular, the immune system attacks the myelin sheath, which acts as an insulator for signal transmission between neurons, and causes a wide range of disabilities. Magnetic resonance imaging (MRI) plays a central role in the diagnosis and management of MS patients because damage to the myelin is visible on MRI. A research question of interest is whether these MRI images can predict MS subtype. Subtype classification is important because disease management and treatment are subtype specific. To answer this question we propose a Bayesian scalar-on-image regression model with scalar outcome (MS subtype) and binary image (presence or absence of lesion at each voxel obtained from MRI) covariates. Parameters of these covariates are spatially varying and are fitted using Gaussian random fields. Scalar covariates such as disease duration are also modeled. Our proposed model is fitted to both simulated data and a real data set consisting of 239 MS patients. A Hamiltonian Monte Carlo (HMC) algorithm is proposed to implement full Bayesian statistical inference. HMC can be more statistically efficient than other Markov Chain Monte Carlo methods when covariates are highly correlated.

email: *cuiguo@umich.edu*

2c. STATISTICAL ESTIMATION OF WHITE MATTER MICROSTRUCTURE FROM CONVENTIONAL MRI

Leah H. Suttner*, University of Pennsylvania

Amanda Mejia, Johns Hopkins School of Public Health

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

Pascal Sati, National Institute of Neurological Disease and Stroke, National Institutes of Health

Daniel S. Reich, National Institute of Neurological Disease and Stroke, National Institutes of Health and Johns Hopkins Bloomberg School of Public Health

Russell T. Shinohara, University of Pennsylvania

Diffusion tensor imaging (DTI) has become the predominant modality for studying white matter integrity in multiple sclerosis (MS) and other neurological disorders. Unfortunately, the use of DTI-based biomarkers in large multi-center studies is hindered by systematic biases which confound the study of disease-related changes. Furthermore, the site-to-site variability in multi-center studies is

significantly higher for DTI than that for conventional MRI-based markers. In our study, we apply the Quantitative MR Estimation Employing Normalization (QuEEN) model to estimate the four DTI measures: MD, FA, RD, and AD. QuEEN uses a voxel-wise generalized additive regression model to relate the normalized intensities of one or more conventional MRI modalities to a quantitative modality, such as DTI. We assess the accuracy of the models by comparing the prediction error of estimated DTI images to the scan-rescan error in subjects with two sets of scans. Across the four DTI measures the performance of models is not consistent: Both MD and RD estimations appear to be quite accurate, while AD estimation is less accurate than MD and RD, and the accuracy of FA estimation is poor. Thus, when only assessing white matter integrity, it is sufficient to acquire conventional MRI sequences alone.

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

2d. SPATIAL STATISTICAL ANALYSIS OF SUICIDAL BEHAVIOR IN HARRIS COUNTY

Aron M. Trevino* University of Texas Health Science Center, San Antonio

Dejian Lai, University of Texas Health Science Center, Houston

In previous studies, the dispersion of suicidal behavior has been found to be spatially clustered. These studies only analyzed the spatial distribution at the state or national level. This study examines the spatial distribution of suicidal behavior at the county level. The spatial distribution was determined using data from the crisis intervention response team of the Harris County Police Department. The techniques used to determine the spatial distribution of suicidal behavior were spatial point pattern analysis and lattice data analysis. For spatial point pattern, analysis the K, G, and F function were computed. For lattice data analysis, the Geary's C and Moran's I functions were computed. The difference of K function was also measured for each year (2011-2014) to see if there was a difference in spatial distribution from one year to the next. Both the lattice data analysis and the spatial point pattern analysis found that suicidal behavior was clustered in Harris County. The K difference functions showed that there was no difference in spatial distribution of suicidal behavior from one year to the next. Future studies can analyze these clusters to see if there is a common trend between them in terms of demographic factors.

email: *trevinoa7@uthscsa.edu*

2e. PENALIZED VARIABLE SELECTION FOR SPATIAL BINARY AND COUNT DATA

Abdhi Amitabha Sarkar*, Michigan State University

Chae Young Lim, Seoul National University

Tapabrata Maiti, Michigan State University

Spatial binary and count data often arise in many scientific applications such as epidemiology and biological sciences. In today's digital world scientists are able to collect large amounts of information. Screening this information has thus become a vital step in statistical analysis. The special characteristic of spatial non-normal data creates versatile challenges. We propose a variable selection technique, that is suitable for binary and count spatially correlated responses which also yields consistent estimates and provide a computational algorithm to ease calculations. The application of this method has been illustrated using a real data example and simulations studies demonstrate the performance of this method.

email: *sarkara1@stt.msu.edu*

2f. RELATING MULTI-SEQUENCE LONGITUDINAL INTENSITY PROFILES AND CLINICAL COVARIATES IN INCIDENT MULTIPLE SCLEROSIS LESIONS

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

Russell T. Shinohara, University of Pennsylvania

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

Matthew K. Schindler, National Institute of Neurological Disease and Stroke, National Institutes of Health

John Muschelli, Johns Hopkins Bloomberg School of Public Health

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

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

Ani Eloyan, Brown University

The formation of multiple sclerosis (MS) lesions is a complex process involving inflammation, tissue damage, and tissue repair - all of which are visible on magnetic resonance imaging (MRI). We introduce a principal component analysis (PCA) and regression model for relating voxel-level, longitudinal, multi-sequence MRI intensities within MS lesions to clinical information. We first characterize the lesion repair process on structural MRI as voxel-level intensity profiles. We perform PCA on the intensity profiles to develop a voxel-level biomarker for identifying slow and persistent, long-term intensity changes within lesion tissue voxels. The biomarker's ability to identify such effects is validated by two experienced clinicians. On a scale of 1 to 4, with 4 being the highest quality, the neuroradiologist gave the biomarker a median quality rating of 4 (95% CI: [4,4]), and the neurologist gave the biomarker a median rating of 3 (95% CI: [3,3]). We then relate the biomarker to the clinical information in a mixed model framework. Treatment with disease-modifying therapies

($p < 0.01$), steroids ($p < 0.01$), and being closer to the boundary of abnormal signal intensity ($p < 0.01$) are all associated with a return of a voxel to intensity values near that of normal-appearing tissue.

email: *emsweene@jhsph.edu*

2g. A BAYESIAN ZERO-INFLATED MULTIVARIATE POISSON MODEL FOR IDENTIFYING FUNCTIONAL CO-ACTIVATION PATTERNS

Caprichia Jeffers*, Emory University

Jian Kang, University of Michigan

Meta-analysis of functional neuroimaging data has become increasingly important recently. Much attention has been paid to detect consistent activation regions or locations across independently performed studies, while limited works have focused on co-activation pattern identifications. To fill this gap, Xue et al. (2014) has proposed a penalized multivariate Poisson model and developed the associated EM algorithm for model inference. However, this method is invalid when foci is sparsely distributed over the brain. To mitigate this problem, we develop a zero-inflated multivariate Poisson distribution for joint modeling the region-level foci counts. For each region and each region pair, we introduce a latent variable following a mixture of zero-point mass and a Poisson distribution left-truncated at zero. The observed region specific foci counts are assumed to be the summation of the latent variables. For each region pair, the dependence between foci counts comes from the associated latent variable, characterizing the number of foci that are co-activated in both regions. We develop an efficient posterior computation algorithm, producing more accurate estimates on the co-activation patterns and the associated brain network compared to the existing approach. We illustrate our methods via extensive simulation studies and a meta-analysis of functional neuroimaging data for emotion studies.

email: *cjeffe4@emory.edu*

h. SPATIAL APPROACH TO AGE-PERIOD-COHORT MODELS

Pavel Chernyavskiy*, National Cancer Institute, National Institutes of Health

Mark P. Little, National Cancer Institute, National Institutes of Health

Philip S. Rosenberg, National Cancer Institute, National Institutes of Health

Age-period-cohort (APC) models are widely used in epidemiology to analyze vital rates. Traditional estimation approaches have several drawbacks, namely: 1) APC models lack parsimony because they fit a categorical deviation from linear trends for each level of age, period and cohort; 2) the estimates cay be sensitive to cells with very small or 0 counts, which may require post-hoc grouping of the data; 3) rates observed at similar ages, periods, and cohorts are treated as independent, but correlations between them may result in incorrect variance estimates. In this paper, we propose a novel approach to estimation of APC models using spatially-correlated Poisson models. We treat event rates as point-referenced data collected on a grid defined by values of age, period, and cohort, with various underlying distance metrics. We incorporate correlation among proximal observations using a spatial random effect. In diverse examples we show that this parameterization complements standard APC models in that: 1) it is more parsimonious; 2) it has unbiased standard errors; and 3) it can fit better when the data are over-dispersed. The spatial approach can also be adapted to assess goodness of fit by application to a residuals analysis and to impute cells with 0 events.

email: *pavel.chernyavskiy@nih.gov*

2i. MODELING NONSTATIONARITY IN SPACE AND TIME

Lyndsay Shand*, University of Illinois, Urbana-Champaign

Bo Li, University of Illinois, Urbana-Champaign

We propose to model a spatio-temporal random field that has nonstationary covariance structure in both space and time domain by extending the dimension expansion method in Bornnetal (2012). Simulations are conducted for both separable and nonseparable space-time covariance models, and the model is illustrated by Illinois wind speed data. Both simulation and data analysis show that by modeling nonstationarity in both space and time improves the predictive performance over stationary covariance models or the model that is nonstationary in space but stationary in time.

email: *lshand2@illinois.edu*

2j. STATISTICAL ANALYSIS OF TRAJECTORIES ON RIEMANNIAN MANIFOLDS

Jingyong Su*, Texas Tech University

In this research we propose to develop a comprehensive framework for registration and analysis of manifold-valued processes. Functional data analysis in Euclidean spaces has been explored extensively in literature. But we study a different problem in the sense that functions to be studied take values on nonlinear manifolds, rather than in vector spaces. Manifold-valued data appear frequently in shape and image analysis, computer vision, biomechanics and many others. If the data were contained in Euclidean space, one would use standard Euclidean techniques and there has been a vast literature on these topics. However, the non-linearity of the manifolds requires development of new methodologies suitable for analysis of manifold-valued data. We propose a comprehensive

framework for joint registration and analysis of multiple manifold-valued processes. The goals are to take temporal variability into account, derive a rate-invariant metric and generate statistical summaries (sample mean, covariance etc.), which can be further used for registering and modeling multiple trajectories.

email: *jingyong.su@ttu.edu*

3. POSTERS: CLINICAL TRIALS, ADAPTIVE DESIGNS AND APPLICATIONS

3a. INVITED POSTER: C-LEARNING: A NEW CLASSIFICATION FRAMEWORK TO ESTIMATE OPTIMAL DYNAMIC TREATMENT REGIMES

Baqun Zhang, Renmin University

Min Zhang*, University of Michigan

Personalizing treatment to accommodate patient heterogeneity and the evolving nature of a disease over time has received considerable attention lately. A dynamic treatment regime is a set of decision rules, each corresponding to a decision point, that determine that next treatment based on each individual's own available characteristics and treatment history up to that point. We show that identifying the optimal dynamic treatment regime can be recast as a sequential classification problem and is equivalent to sequentially minimizing a weighted expected misclassification error. This general classification perspective targets the exact goal of optimally individualizing treatments and is new and fundamentally different from existing methods. Based on this fresh classification perspective, we propose a novel, powerful and flexible C-learning algorithm to learn the optimal dynamic treatment regimes backward sequentially from the last stage till the first stage. C-learning is a direct optimization method that directly targets optimizing decision rules by exploiting powerful optimization/classification techniques and it allows incorporation of patient's characteristics and treatment history to dramatically improves performance, hence enjoying the advantages of both the traditional outcome regression based methods (Q-and A-learning) and the more recent direct optimization methods. The superior performance and flexibility of the proposed methods are illustrated through extensive simulation studies.

email: *mzhangst@umich.edu*

3b. AN EXTENSION OF THE CLOSURE PRINCIPLE FOR THE IDENTIFICATION OF INDIVIDUAL EFFICACIOUS ENDPOINTS WHEN USING COMPOSITE ENDPOINTS IN CLINICAL TRIALS

Jaclyn A. McTague*, Prosoft Clinical

Dror Rom, Prosoft Clinical

Chen Chen, Prosoft Clinical

Composite endpoints are used when several therapeutic effects of a drug are combined to provide a more powerful overall signal of the drug effect instead of the assessment of individual endpoints. This is done when each individual therapeutic effect may be too small to detect even in a large clinical study, while a combination of several endpoints may lead to a stronger signal. The downside of composite endpoints is their inability to provide a more concise understanding of which, if any, of the individual effects can be singled out as a major contributor to the overall effect, unless a more detailed analysis is done. In this presentation, we devise a method of conducting inferences on individual endpoints using a simple extension of the closure principle. The method is shown to provide strong control of the Familywise Error Rate (FWER). Examples from clinical trials are provided.

email: *J.McTague@ProsoftClinical.com*

3c. A LIKELIHOOD DESIGN FOR SINGLE ARM PHASE II GROUP SEQUENTIAL CLINICAL TRIALS WITH TIME-TO-EVENT ENDPOINTS

Wei Wei*, Medical University of South Carolina

Elizabeth Garrett-Mayer, Medical University of South Carolina

A phase II oncology trial generally takes the form of single-arm two-stage design and is often based on a binary endpoint such as tumor response. Tumor response is not always a good surrogate for time-to-event (TTE) endpoints and in some cases a surrogate endpoint is not necessary because overall survival is available within a relatively short period of time. Frequentist and Bayesian designs have been developed for phase II trials with TTE endpoints, but these designs rely on parametric assumptions about data yet to be collected. This paper proposes a likelihood design as an alternative to Bayesian approach in single arm phase II trials with TTE endpoints. Our design is based on empirical likelihood which is a non-parametric analogue of the likelihood function. Compared to the Bayesian design, the EL method is much easier to implement because there is no need to worry about choosing inappropriate priors. We showed the empirical likelihood-based design can achieve a desired PET in a variety of simulation scenarios. In some cases, the EL design outperforms the exponential inverse gamma model based Bayesian design. Motivating examples from a trial of Hepatocellular Carcinoma were provided to illustrate the use of empirical likelihood method.

email: *weiwe@musc.edu*

3d. COMPARING FOUR DOSE ESCALATION DESIGNS IN PHASE I ONCOLOGY TRIALS

Zhao Yang*, University of Southern California and Biometrics, Medivation, Inc.

Rui Li, Biometrics, Medivation, Inc.

Suman Bhattacharya, Biometrics, Medivation, Inc.

In Phase I oncology trials, the traditional 3+3 design and its variants are often unable to determine the Maximum Tolerated Dose (MTD) optimally. As a result, there were a number of alternative designs, in which the Continual Reassessment Method (CRM), the Bayesian Logistic Regression Model (BLRM) and the modified Toxicity Probability Interval (mTPI) method are often cited. In practice, it is important to have a thorough understanding of these alternative designs and their relative performance in order to make an optimal choice. In this study, we present a comprehensive comparison of the performance of four dose escalation designs in Phase I oncology trials, 3+3 H, CRM, BLRM and mTPI, by looking at their operating characteristics in simulation studies in five scenarios with different true DLT rates of each dose level. In addition, we use interim monitoring analysis to compare escalation decisions made by each of them given the same data. In both simulation studies and interim monitoring analysis, the 3+3 H method turns out to be the most conservative in nature and tends to underestimate MTD. On the other hand, CRM tends to overestimate MTD, mTPI performs better than 3+3 H and BLRM performs best in general.

email: *yang19@usc.edu*

3e. NON-INFERIORITY STUDIES WITH MULTIPLE REFERENCE TREATMENTS AND HETEROGENEOUS VARIANCES

Li-Ching Huang*, Vanderbilt University

Miin-Jye Wen, National Cheng-Kung University, Taiwan

Yu Shyr, Vanderbilt University

Non-inferiority (NI) studies are gaining popularity in clinical trials. The objective of NI trials is to explore potential substitutes (new treatments) for existing treatments (reference treatments) by verifying that new treatments maintain a sizable portion of the efficacy of reference treatments. The loss of efficacy (NI margin) of new treatments, as compared to reference treatments, can be compensated by its other benefits, such as alleviating side effects, lowering costs, and simplifying intricate treatment regimens. Statistical methods have been developed for the simultaneous testing of NI of multiple new treatments against multiple reference treatments. However, these procedures are based upon the assumption of equal variances in all treatments. In this paper, we use a simulation study to explore the undesirable effects of a violation of the homogeneous variance assumption of using these procedures on the familywise Type I error rate (FWE). To remedy the problem, we propose procedures that are more appropriate because these procedures are able to control the FWE. A power study is then conducted to compare the different procedures, and a clinical example is given for illustrative purposes.

email: *li-ching.huang@vanderbilt.edu*

3f. BAYESIAN CLINICAL TRIAL DESIGN FOR A VALIDATION STUDY OF MOLECULAR ALTERATION IDENTIFICATION

Xiaoxiao Lu*, West Virginia University

Sijin Wen, West Virginia University

Identifying the molecular alterations that can distinguish any particular cancer cell from a normal cell will eventually help to define the phenotype of cancer cells, and predict their pathologic behavior including responsiveness to treatment. To date, a number of diagnosis methods have been developed aiming at precise detection, diagnosis and prevention of cancer. Validation of different diagnosis methods is a high priority. We propose a Bayesian clinical trial design to compare two diagnosis methods for molecular alteration identification and to examine the equivalence of these two methods. In particular, we assess the posterior probability of finding the molecular alterations based on the agreement of two methods. This probability can be used for validation of diagnosis methods and for justification of sample size with early stopping rules. We provide general guidelines for application and illustrate the trial design with different scenarios of prevalence, type I and type II errors in the diagnosis methods.

email: *siwen@hsc.wvu.edu*

3g. AUC REGRESSION FOR MULTIPLE COMPARISONS TO A CONTROL WITH APPLICATION IN DETERMINING THE MINIMUM EFFECTIVE DOSE

Johanna S. Van Zyl*, Baylor University

Jack D. Tubbs, Baylor University

The relationship between the non parametric AUC and Mann-Whitney statistic has been determined by Bamber (1975). Further, Pepe et al. (2003) introduce a semi parametric generalized linear model framework to adjust the AUC for covariates. Thus, we can adjust the Mann-Whitney statistic for covariates by modeling the AUC. Zhang et al. (2011) correct the standard errors of the regression coefficients using a combination of Delong's method (1988) and the Delta method. Buros et al. (2015) extend the concepts of the generalized linear modeling framework for the AUC to the Jonckheere Terpstra (JT) test such that we can adjust the JT test for discrete covariates. Buros et al. further explores multiple comparisons with discrete covariates assuming an ordered alternative. We extend this research to multiple comparisons to a control for an ordered alternative with applications to determining the minimum effective dose. We evaluate the method in a Monte Carlo study with a control arm and three treatment arms. Further, we demonstrate the method with a synthetic data set based on summary statistics from a real clinical trial.

email: *Johanna_Van_Zyl@baylor.edu*

3h. BAYESIAN APPROACH TO SAMPLE SIZE DETERMINATION FOR MULTILEVEL LOGISTIC REGRESSION MODELS WITH MISCLASSIFIED OUTCOMES

Tyler W. Nelson*, Baylor University

James D. Nelson, Baylor University

We develop a simulation based procedure for determining the required sample size in a multi-level logistic regression model when response data are subject to misclassification. A Bayesian average power criterion, proposed is used to determine a sample size that provides probability of correctly selecting the direction of an association between predictor variables and the probability of event occurrence. Though the model we consider is logistic, the methods proposed could also be used for other count models such as the Poisson. We consider two scenarios, one in which we only have one diagnostic test with misclassification of the response variables, and the second where we will have two independent diagnostic tests in which the response variables are misclassified. Finally we compare the Bayesian power of both scenarios using a simulation approach.

email: *tyler_nelson@baylor.edu*

3i. AN EVALUATION OF CONSTRAINED RANDOMIZATION FOR THE DESIGN AND ANALYSIS OF GROUP-RANDOMIZED TRIALS

Fan Li*, Duke University

Yuliya Lokhnygina, Duke University

David Murray, National Institutes of Health, Office of Disease Prevention

Patrick Heagerty, University of Washington

Elizabeth DeLong, Duke University

In group-randomized trials, a frequent practical limitation to adopting rigorous research designs is that only a small number of groups may be available, and therefore simple randomization cannot be relied upon to balance key group-level prognostic factors. Constrained randomization is an allocation technique proposed for ensuring balance, and can be used together with a permutation test for randomization-based inference. However, several statistical issues have not been thoroughly studied under constrained randomization. Therefore, we evaluated several key issues including: impact of candidate set size and balance metric used to guide randomization; choice of adjusted versus unadjusted analysis, and model-based versus randomization-based tests. We conducted simulation studies to compare the type I error and power of the F-test and the permutation test in the presence of group-level potential confounders. Our results indicate that the adjusted F-test and the permutation test perform similarly in terms of power regardless of randomization designs. Under constrained randomization, however, the unadjusted F-test is conservative while the unadjusted permutation test carries the desired type I error rate as long as the candidate set size is not too small; the unadjusted permutation test is consistently more powerful than the unadjusted F-test, and gains power as candidate set size changes. Finally, we caution against the inappropriate specification of permutation distribution under constrained randomization.

email: *frank.li@duke.edu*

3j. OPTIMAL GROUP SEQUENTIAL DESIGN

Qi An*, University of Florida

A group sequential design if referred to an optimal one if the expected sample size under specified alternative is minimized for a given choice of significance level and power among tests with a given sequence of groups of specific sizes. We concern optimal two-sided tests with acceptance region of the null hypothesis at early stages with and without known variance. We apply simulation method to solve multi-dimensional integration problem in calculation.

email: *anqi@ufl.edu*

3k. AN INFORMATIVE PRIOR APPROACH TO A BIVARIATE ZERO-INFLATED POISSON REGRESSION MODEL

Madeline L. Drevets*, Baylor University

John W. Seaman, Baylor University

Bivariate zero-inflated poisson (BZIP) regression models have been used in several applications to model bivariate count data with excess zeros. Bayesian treatments of BZIP models have focused on diffuse prior structures for model parameters. These parameters depend on covariates through canonical link, generalized linear models. A common, relatively noninformative Bayesian prior approach is to place diffuse priors on regression coefficients and subsequently induce priors on the model parameters of interest. However, such an approach may be problematic in some cases as it can result in identifiability issues with the estimation of some parameters. We present an example to illustrate this. We offer an informative prior approach for BZIP model parameters. In particular, we propose a prior structure that uses expert opinion to elicit informative priors. Finally, we present an example in the medical context to illustrate our methods.

email: *Madeline_Drevets@baylor.edu*

3l. SENSITIVITY IN PRIOR ELICITATION

Somer E. Blair*, Baylor University

David Kahle, Baylor University

John W. Seaman, Jr., Baylor University

Prior elicitation refers to the process of converting expert opinion into a probability distribution for use in a Bayesian data analysis. The most common method of elicitation has the statistician obtain

a collection of distribution summaries from the expert (mean, mode, percentiles, etc.) and then convert these into the standard parameters of an assumed family (e.g. the shape parameters of a beta distribution). One question that arises in this process is the sensitivity of the resulting parameters, and consequently the prior, to slight deviations in the expert's specifications, which may arise from his inability to precisely quantify his beliefs, even for summaries chosen to be amenable to specification. The hope is that small deviations from the true values (correct expert beliefs) should not make substantial differences in the ensuing prior or subsequent analysis, but this hope is not always realized. In this presentation we consider ways to measure the sensitivity of prior specifications on imprecisely specified summaries and demonstrate that, depending on the type of specification, small mis-specifications can result in dramatically different priors.

email: *somer_blair@baylor.edu*

3m. COMBINING NON-RANDOMIZED AND RANDOMIZED DATA IN CLINICAL TRIALS USING COMMENSURATE PRIORS

Hong Zhao*, University of Minnesota

Brian P. Hobbs, University of Texas MD Anderson Cancer Center

Haijun Ma, Amgen Inc.

Qi Jiang, Amgen Inc.

Bradley P. Carlin, University of Minnesota

Randomization eliminates selection bias, and attenuates imbalance among study arms with respect to prognostic factors. Thus, information from randomized clinical trials (RCTs) is typically considered the gold standard for comparing therapeutic interventions in confirmatory studies. However, RCTs are limited in study population due to strict inclusion criteria. By contrast, observational studies (OSs) reflect a broader patient population. However, OSs often suffer from selection bias, and may yield invalid treatment comparisons. Therefore, combining information from OSs with data from RCTs is often criticized due to the limitations of OSs. In this article, we combine randomized and non-randomized substudy data from FIRST, a recent HIV drug trial. We develop hierarchical Bayesian approaches devised to combine data from all sources simultaneously while explicitly accounting for potential discrepancies in the sources' designs. Specifically, we describe a two-step approach combining propensity score matching and Bayesian hierarchical modeling to integrate information from non-randomized studies with data from RCTs, to an extent that depends on the estimated commensurability of the data sources. We investigate our procedure's operating characteristics via simulation. Our findings elucidate the extent to which well-designed non-randomized studies can complement RCTs.

email: *zhao0504@umn.edu*

4. POSTERS: SURVIVAL ANALYSIS

4a. QUANTILE RESIDUAL LIFE REGRESSION WITH LONGITUDINAL BIOMARKER MEASUREMENTS FOR DYNAMIC PREDICTION

Ruosha Li*, University of Texas Health Science Center, Houston

Xuelin Huang, University of Texas MD Anderson Cancer Center

Jorge Cortes, University of Texas MD Anderson Cancer Center

Residual life is of great interest to patients with life-threatening disease. It is also important for clinicians who estimate prognosis and make treatment decisions. Quantile residual life has emerged as a useful summary measure of the residual life. It has many desirable features, such as robustness and easy interpretation. In many situations, the longitudinally collected biomarkers during patients' follow-up visits carry important prognostic value. In this work, we study quantile regression methods that allow for dynamic predictions of the quantile residual life, by flexibly accommodating the post-baseline biomarker measurements in addition to the baseline covariates. We propose unbiased estimating equations that can be solved via existing L-1 minimization algorithms. The resulting estimators have desirable asymptotic properties and satisfactory finite-sample performance. We apply our method to a study of chronic myeloid leukemia to demonstrate its usefulness as a dynamic prediction tool.

email: *ruosha.li@uth.tmc.edu*

4b. EVALUATING USE OF A COX REGRESSION MODEL IN LANDMARK ANALYSIS TO APPROXIMATE AN ILLNESS-DEATH MODEL

Krithika Suresh*, University of Michigan

Jeremy M.G. Taylor, University of Michigan

Alex Tsodikov, University of Michigan

The landmarking approach to dynamic prediction incorporates time-dependent information accrued during follow-up to improve survival prediction probabilities. For several chosen landmark times, a dataset is created containing patients still at risk for failure and their covariate values at that time. These datasets are stacked and a simple Cox model is fit to the residual time. To account for overlap between the datasets, the Cox model coefficients are restricted to have a parametric form that is a function of the landmark time. Our goal is to determine whether a Cox model is appropriate to achieve consistency between the predictions at the different landmark times. Considering an illness-death model, we wish to find the form of the corresponding residual time model in a landmarking approach and assess whether it is consistent with

a Cox model. By equating the residual time distributions under both approaches we identify the structure of the Cox model baseline hazard and covariate effects that corresponds to the landmark illness-death model. In the landmark setting, the covariate effects should be independent of the residual time. Since a simple Cox regression does not satisfy this condition, we explore alternative flexible model structures that provide a consistent and valid approximation.

email: *ksuresh@umich.edu*

4c. ESTIMATING ENVIRONMENTAL MODIFICATION ON COEFFICIENTS OF COX PROPORTIONAL HAZARDS MODEL IN THE STUDY OF SEXUAL MATURATION

Huazhen Lin, Southwestern University of Finance and Economics

Peter Song, University of Michigan

Ling Zhou*, University of Michigan

Phthalates and bisphenol A (BPA), known as potential endocrine disrupters, are widely used production chemicals. Pregnant women are easily exposed to these toxic agents, and their effects on child growth and development are of great interest in public health. In this project, we aim to evaluate the adverse impact of prenatal exposures on sexual maturation. Based on the data from the Early Life Exposure in Mexico to Environmental Toxicants (ELEMENT) cohorts, we investigate if, and how prenatal exposure to phthalates and BPA may modify the predictive relationship of growth characteristics with timing of puberty in girls. We propose a new type of Cox proportional hazards (PH) model with varying index coefficients, which enables evaluation of nonlinear interaction effects between mixtures of toxic agents on the time to sexual maturation. We develop a global partial likelihood method based on a new local smoothing technique. We establish some key large-sample properties, including estimation consistency, asymptotic normality and semi-parametric efficiency. Both proposed model and estimation methods are illustrated by extensive simulation studies and applied to the analysis of an ELEMENT data consisting of 113 girls aged 8.1 to 13.7 years.

email: *zholing@umich.edu*

4d. INFERENCE OF TRANSITION PROBABILITIES IN MULTI-STATE MODELS USING ADAPTIVE INVERSE PROBABILITY CENSORING WEIGHTING TECHNIQUE

Ying Zhang*, Medical College of Wisconsin

Meijie Zhang, Medical College of Wisconsin

Inverse probability censoring weighting (IPCW) technique has been used extensively for analyzing right censored time to event data. In multistate modeling, censoring distribution has been estimated by a simple Kaplan-Meier estimator, where individuals in any one of the transient states at the end of the study were considered as censored individual. Our simulation study shows that such simple IPCW estimate may lead to biased estimates when analyzing multi-level complicated multistate models. We propose a state-dependent adaptive IPCW (AIPCW) technique for estimating and modeling transition probabilities. Proposed AIPCW estimates are asymptotically unbiased. We conduct a simulation study to show that proposed AIPCW method performs well. We apply the method to a bone marrow transplant data to estimate the transition probabilities, and to assess the graft-verse-host disease (GVHD) effect on survival outcomes.

email: *yizhang@mcw.edu*

4e. MEASURING THE EFFECTS OF A TIME-DEPENDENT TREATMENT ON CORRELATED RECURRENT AND TERMINAL EVENTS USING FRAILTY-BASED PROGNOSTIC MODELS

Abigail R. Smith*, University of Michigan

Douglas E. Schaubel, University of Michigan

In many observational studies, treatment is time-dependent (i.e., assigned after time 0) and related to prognostic factors including history of a recurrent event process such as hospitalization. This event process in turn is usually related to a terminating event process such as death that stops all subsequent realizations of the recurrent event. We propose a two-stage method to determine the association between a time-dependent treatment and both recurrent and terminal events using sequential stratification and frailty prognostic models to mimic a randomized experiment. In the first stage, the treatment-free recurrent and terminal event processes are modeled among all patients using frailty models in order to determine trajectories in the absence of treatment. Patients that receive treatment are then matched to other treatment-eligible subjects with similar trajectories to form strata. Stratified models are then fit for the observed recurrent and terminal events to estimate treatment associations. A variation of inverse probability of censoring weighting is used to account for the dependent censoring of matched patients that are subsequently treated. The method is tested in moderate sized samples through simulation, and an application to organ transplant data is presented.

email: *abbysmit@umich.edu*

4f. SEMIPARAMETRIC BAYESIAN ESTIMATION OF QUANTILE FUNCTION FOR SURVIVAL DATA WITH CURED FRACTION

Cherry C.H. Gupta*, Florida State University

Juliana Cobre, Universidade de São Paulo

Andriano Polpo, Universidade Federal de São Carlos

Debjayoti Sinha, Florida State University

Existing cure rate survival models are generally not convenient for expressing and estimating the survival quantiles of a patient with specified covariate values. We propose a novel class of cure rate

model, the transform both sides cure rate model (TBSCRM), which can be used to make inference about both the cure rate and the survival quantiles. We develop the Bayesian inference about the covariate effects on the cure rate as well as on the survival quantiles via Markov Chain Monte Carlo (MCMC) tools. We also show that in our simulation studies and application to the breast cancer survival data from the National Cancer Institute's Surveillance, Epidemiology and End Results (SEER) database, the TBSCRM based Bayesian method outperforms existing cure rate models based methods.

email: *cgupta@stat.fsu.edu*

4g. DYNAMIC PROGNOSIS TOOL OF ACUTE GRAFT-VERSUS-HOST DISEASE BASED ON BIOMARKERS

Yumeng Li*, University of Michigan

Thomas 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 Landmark analysis. We consider settings in which the biomarkers are subject to measurement errors or not, and sample size is also a key factor in best approach choice. 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*

4h. PROPORTIONAL HAZARDS MODEL WITH A CHANGE POINT FOR CLUSTERED EVENT DATA

Yu Deng*, University of North Carolina, Chapel Hill

Donglin Zeng, University of North Carolina, Chapel Hill

Jinying Zhao, Tulane University

Jianwen Cai, University of North Carolina, Chapel Hill

In many epidemiology studies, family data with survival endpoints are collected to investigate the association between risk factors and disease incidence. Sometimes the risk of the disease may change when a certain risk factor exceeds a certain threshold. Finding this threshold value could be important for disease risk prediction and diseases prevention. In this work, we propose a change-point proportional hazards model for clustered event data. The model incorporates the unknown threshold of a continuous variable as a change point in the regression. The marginal pseudo-partial likelihood functions are maximized for estimating the regression coefficients and the unknown change point. We develop a supremum test based on robust score statistics to test the existence of the change point. The inference for the change point estimator is based on the m out of n bootstrap. We establish the consistency and asymptotic distributions of the proposed estimators. The finite-sample performance of the proposed method is demonstrated via extensive simulation studies. Finally, the Strong Heart Family Study dataset is analyzed to illustrate the methods.

email: *yudeng@live.unc.edu*

4i. A CLASS OF TWO-SAMPLE TESTS FOR QUANTILE RESIDUAL LIFE TIME

Yimeng Liu*, University of Pittsburgh

Abdus S. Wahed, University of Pittsburgh

Quantile residual lifetime (QRL) is of significant interest in many clinical studies as an easily interpretable quantity compared to other summary measures of survival distributions. In cancer or other fatal diseases, often treatments are compared based on the distributions or quantiles of the residual lifetime. Thus a common question arises: how to test the equality of the QRL between two populations. In this paper, we propose two classes of tests to compare two QRLs: one class is based on the difference between two estimated QRLs, and the other based is on the estimating functions of the QRL, where estimated QRL from one sample is plugged into the QRL-estimating-function of the other sample. We outline the asymptotic properties of these test statistics. Simulation studies demonstrate that proposed tests produced type I errors closer to the nominal level and are superior to some existing tests based on both type I error and power. Our proposed statistics are also computationally less intensive and more straightforward to be used compared to tests based on the confidence intervals.

email: *yil103@pitt.edu*

4j. SEMIPARAMETRIC REGRESSION ANALYSIS OF INTERVAL-CENSORED COMPETING RISKS DATA

Lu Mao*, University of North Carolina, Chapel Hill

Danyu Lin, University of North Carolina, Chapel Hill

Donglin Zeng, University of North Carolina, Chapel Hill

In clinical and epidemiological studies, competing risks data arise when the subject can experience one, and only one, of several mutually exclusive types of events. Competing risks data are often subject to interval censoring when the events are asymptomatic and can only be detected by periodic examinations. The presence of multiple distinct types of events and the lack of exact observation times pose serious challenges for the analysis of interval-censored competing risks data. In this paper, we propose a general

class of semiparametric transformation regression models for the cumulative incidence function of competing risks, which incorporates the proportional hazards and proportional odds models as special cases. We allow covariates to be time-dependent and accommodate missing event type information. We develop a novel EM algorithm to compute the nonparametric maximum likelihood estimators (NPMLEs), and establish the consistency, asymptotic normality, and semiparametric efficiency of the NPMLEs. Extensive numerical studies show that our methods perform well in finite samples. A well-known HIV/AIDS study is provided to illustrate our methods.

email: *lmao@unc.edu*

4k. SEMIPARAMETRIC REGRESSION MODEL FOR RECURRENT BACTERIAL INFECTIONS AFTER HEMATOPOIETIC STEM CELL TRANSPLANTATION

Chi Hyun Lee*, University of Texas MD Anderson Cancer Center

Xianghua Luo, University of Minnesota

Chiung-Yu Huang, Johns Hopkins University

Todd E. DeFor, University of Minnesota

Claudio G. Brunstein, University of Minnesota

Daniel J. Weisdorf, University of Minnesota

Patients who undergo hematopoietic stem cell transplantation (HSCT) often experience multiple bacterial infections during the early post-transplant period. The interoccurrence times or gap times between recurrent infections after transplant are the natural outcome of interest. In this study, we focus on modeling the relationship between inter-infection gap times and patient- and transplant-related risk factors. Conventional survival models such as the Cox proportional hazards model or the accelerated failure time model for univariate survival data are not directly applicable to recurrent gap time data due to the well-known induced dependent censoring problem caused by within-subject correlation among gap times. Despite rich literature on recurrent gap time data, existing semiparametric methods commonly assume that all gap times are identically distributed. Hence these methods are not applicable to our post-transplant infection data because the initial event of our data is transplant, which is a different type of event than the recurrent events. Therefore, a method which allows the time from transplant to the first infection to distribute differently than the gap times between consecutive infections is necessary. In this article, we propose a semiparametric regression model to evaluate the covariate effects on time from transplant to the first infection and on gap times between consecutive infections simultaneously. An application of the proposed method to a post-HSCT bacterial dataset collected at the University of Minnesota is presented.

email: *leex5865@umn.edu*

4l. ESTIMATION AND MODELING OF SEXUAL PARTNERSHIP DATA

Yared Gurmu*, Harvard University

Data that describe sexual partnership duration are useful for modeling spread of sexually transmitted infections. Such data are commonly obtained through surveys that collect information on relationships that are ongoing during a fixed time window. This sampling mechanism leads to duration data that are left truncated and right censored; and have been analyzed using the standard truncation product limit estimator (TPLE). In this presentation, we will first describe a common sampling scheme for collecting sexual partnership data and investigate conditions for unbiased estimation of the duration distribution. We will then propose a stochastic expectation maximization algorithm (stEM) coupled with rejection-sampling scheme in order to estimate transition rates from a state of celibacy to monogamy and to concurrency (or vice versa). In particular, this paper will address maximum likelihood estimation via stEM when our observed data includes information on the number of certain types of transitions without specifying the sojourn time in the states. For example, with regards to partnership data, the total life time number of partnerships maybe known even though the sojourn time of each of the partnerships in the different states may not be known. Simulation results showing the performance of the stEM will be presented. Simple model validation strategies based on a statistic from the stochastic EM will be presented. Lastly, we will provide an application example based on partnership data collected from a clinic cohort in Kwazulu-Natal, South Africa.

email: *yared@mail.harvard.edu*

4m. A THRESHOLD-FREE PROSPECTIVE PREDICTION ACCURACY MEASURE FOR CENSORED TIME TO EVENT DATA

Yan Yuan*, University of Alberta

Bingying Li, Simon Fraser University

Qian Zhou, Simon Fraser University

Prediction performance of a risk scoring system needs to be evaluated before the adoption of the risk score system in clinical practices. In risk prediction, the primary interest is to predict the time-dependent binary event status at a (set of) pre-specified future time t0. A larger risk score should be associated with a higher risk of developing the event by time t0. Thus, the natural performance metrics for risk scoring systems are prospective accuracy measures – such as the positive and negative predictive values (PPV(t0), NPV(t0)). However, these two measures are only defined for dichotomous scores, which necessitate the use of a subjective cut-off threshold c to dichotomize risk scores which are typically continuous or ordinal. We propose a threshold-free metric, average positive predictive value (AP(t0)), averaging PPV(t0) over true positive frac-

tion (TPF(t0)) for the entire range of the risk scores. The definition and interpretation of AP(t0) will be given. We conduct a simulation study to examine the finite sample performance of the proposed nonparametric estimator and inference procedure for AP(t0). Lastly, we illustrate this metric on a real data example, comparing two risk score systems for predicting heart failure.

5. POSTERS: CAUSAL INFERENCE

5a. INVITED POSTER: ESTIMATING CAUSAL EFFECTS OF POWER PLANT REGULATIONS: BIPARTITE CAUSAL INFERENCE WITH INTERFERENCE

Corwin M. Zigler*, Harvard School of Public Health

Chanmin Kim, Harvard School of Public Health

A vast array of air pollution research motivates various regulatory programs that compel US power plants to reduce harmful pollution emissions. Increasingly contentions legal, legislative, and political pressures present the urgent need to evaluate the health benefits of such policies, but doing so is met with methodological challenges that are not accommodated by existing causal inference literature. A defining feature of power plant regulatory policies is that air pollution travels across space, meaning that health outcomes at a given location are determined not by a single action, but by collections of actions taken at multiple power plants. As a consequence, evaluating these actions is met with two key challenges. First, interventions are implemented at power plants, but key questions for regulatory policy pertain to how emissions reductions unfold throughout the atmosphere to affect pollution and health outcomes across the country. Thus, the units at which the interventions are defined and implemented (power plants) differ from the units at which outcomes are defined and measured (residential locations or individuals). We term this setting one of bipartite causal inference. Second, pollution exposure and health outcomes at a given location are dependent upon interventions applied at many power plants, which is known in the causal inference literature as interference. Collectively, we term the setting of causal inference with interference among two levels of observational unit one of bipartite causal inference with interference. We develop new methods for estimating causal effects in this context and illustrate the methods by evaluating causal effects of strategies to control harmful pollution emissions from coal-fired power plants across the US.

email: *czigler@hsph.harvard.edu*

5b. ON JUSTIFYING THE USE OF SUMMARY COMORBIDITY MEASURES FOR HEALTH SERVICES RESEARCH

Elizabeth A. Gilbert*, Temple University

Robert T. Krafty, Temple University and University of Pittsburgh

Richard J. Bleicher, Fox Chase Cancer Center

Brian L. Egleston, Fox Chase Cancer Center

Prognostic scores have been proposed as outcome based confounder adjustment scores akin to propensity scores. However, prognostic scores have not been widely used in the substantive literature. Instead, comorbidity scores, which are limited versions of prognostic scores, have been used extensively by clinical and health services researchers. A comorbidity is an existing disease an individual has in addition to a primary condition of interest, such as cancer. Comorbidity scores are used to reduce the dimension of a vector of comorbidity variables into a single scalar variable. Such scores are often added to regression models with other non-comorbidity variables such as patient demographics for prognostic and confounder adjustment purposes. Despite their widespread use, the properties of and conditions under which comorbidity scores are valid dimension reduction tools in statistical models is largely unknown. We expand on our previous work to show that under relatively standard assumptions within a causal inference framework, comorbidity scores can have equal prognostic and confounder-adjustment abilities as the individual comorbidity variables. The impact of using comorbidity scores during treatment effect estimation, particularly under interaction assumptions, is also examined. A breast cancer example using the SEER Medicare Database is provided.

email: *elizabeth.gilbert@temple.edu*

5c. NONPARAMETRIC ESTIMATION OF COMPLIER EFFECTS WITH CONTINUOUS INSTRUMENTAL VARIABLES

Edward H. Kennedy*, University of Pennsylvania

Dylan S. Small, University of Pennsylvania

WITHDRAWN

Instrumental variables are very commonly used to estimate effects of treatments that are afflicted by unmeasured confounding. Continuous instruments are very popular in practice, and include for example measures of distance or physician preference. However, previous approaches for continuous instruments require parametric assumptions for both estimation and identification. In this work we develop novel semiparametric approaches for nonparametrically identified effects. Specifically we propose non- and semi-parametric estimators of effects among compliers at given instrument values, i.e., people who would take treatment at one instrument value but not at another. We also construct estimators for stochastic complier effects, i.e., effects among people who would only

take treatment had their observed instrument value been increased but not decreased by some amount. We study asymptotic properties of our proposed estimators, and use them to study the effect on mortality for premature babies born at hospitals whose neonatal intensive care units had high versus low technical capacity, using travel time as an instrument.

email: edwardh.kennedy@gmail.com

5d. PROPENSITY SCORE MATCHING FOR CLUSTERED DATA

Mi-Ok Kim, Cincinnati Children's Hospital Medical Center
Bo Lu, The Ohio State University
Yu Wang*, Cincinnati Children's Hospital Medical Center
Chunyan Liu, Cincinnati Children's Hospital Medical Center
Edward Nehus, Cincinnati Children's Hospital Medical Center
Maurizio Macaluso, Cincinnati Children's Hospital Medical Center

Matching is a popular approach to reducing confounding in observational studies. Various matching techniques/designs are available. Most of them were developed for causal inference with independent data, and little is known about their application to clustered data. We consider two contrasting matching techniques, optimal full matching (Rosenbaum 1991; Hansen, 2004) and 1:1 nearest neighbor matching, and investigate different implementation strategies in the hierarchical clustered setting. Optimal full matching utilizes all available samples, whereas 1:1 nearest neighbor matching may lead to reduced sample sizes. If cluster level heterogeneous effects are of concern, matching shall be restricted to be within clusters, in which case the matching quality might not be good. To produce more closely matched pairs, we could allow matching across clusters. It may introduce bias, however. We propose a hybrid hierarchical matching strategy to achieve a better balance between matching quality and bias consideration. Simulation studies are conducted for comparison, which mimics a real world setting of multi-center kidney transplant cohort study.

email: yu.wang@cchmc.org

5e. MACHINE LEARNING FOR CHARACTERIZATION OF DEVELOPING NEURONAL CULTURES

Diana R. Hall*, Columbia University
Ellese Cotterill, Cambridge University
Kathleen Wallace, United States Environmental Protection Agency
William Mundy, United States Environmental Protection Agency
Stephen J. Eglen, Cambridge University
Timothy J. Shafer, United States Environmental Protection Agency

Information rich biologic assays generate a large number of features to describe underlying biology activity. One such assay is multi-well micro electrode arrays (MEA) platform which record the spiking of cultures of neurons. Machine learning techniques are a useful tool to summarize underlying structure in high dimensional data so as to gain interpretable insights into biologic activity. In the present work, random forest techniques are used to rank features extracted from MEA recordings. The goal of the analysis is to uncover those features of the neuronal networks that best discriminate between different ages of the neuronal cultures. Support vector machine was used to ascertain the extent of the separation between culture ages.

email: dianaransomhall@yahoo.com

5f. GLiDeR: DOUBLY ROBUST ESTIMATION OF CAUSAL TREATMENT EFFECTS WITH THE GROUP LASSO

Brandon Lee D. Koch*, University of Minnesota
David M. Vock, University of Minnesota
Julian Wolfson, University of Minnesota

The efficiency of doubly robust estimators of the average causal effect (ACE) of a treatment can be improved by including in the treatment and outcome models only those covariates which are related to both treatment and outcome (i.e., confounders) or related only to the outcome. However, it is often challenging to identify such covariates among the large number that may be measured in a given study. In this paper, we propose GLiDeR, a novel variable selection technique for identifying confounders and predictors of outcome using an adaptive group lasso approach that simultaneously performs coefficient selection, regularization, and estimation across the treatment and outcome models unlike traditional variable selection methods which consider each model separately. The selected variables and corresponding coefficient estimates are used in a standard doubly robust ACE estimator. We provide asymptotic results and conduct a comprehensive simulation study which shows that GLiDeR is more efficient than doubly robust methods using standard variable selection techniques and has substantial computational advantages over a recently proposed doubly robust Bayesian model averaging method. We illustrate our method to estimate the causal treatment effect of bilateral versus single-lung transplant on forced expiratory volume in one year after transplant using an observational registry.

email: kochx402@umn.edu

5g. THE VALIDATION AVERAGE PREDICTIVE EFFECT (VAPE) FOR EVALUATING RISK PREDICTION TOOLS

Andreas N. Strobl*, Technical University Munich, Germany
Donna P. Ankerst, Technical University Munich, Germany and University of Texas Health Science Center, San Antonio

As risk prediction tools are developed, the need to validate them on external populations becomes important before they can be widely distributed in clinical practice. Our experience with validation of the Prostate Cancer Prevention Trial (PCPT) Risk Calculator on ten international cohorts comprising the Prostate Biopsy Collaborative Group (PBCG) has revealed that validation characteristics of a risk tool, as measured by calibration and discrimination, can vary widely depending on the cohort, even after conditioning on patient and cohort properties. One explanation is differences among cohorts as to who is included in the study, or in other words referred for ascertainment of the disease endpoint. In this talk, we take a causal inference approach to validation, extending the Survival Average Causal Effect (SACE) of Hayden, Pauler, and Schoenfeld (2005) to VAPE for resolving differences between participants referred to disease outcome determination in the training versus testing sets. The VAPE evaluates a validation metric, such as the area-underneath-the-receiver-operating-characteristic-curve (AUC), in the principal stratum of subjects who would have been ascertained for the disease outcome in both the training set used to build the risk prediction model and the testing set used to validate it.

email: *a.strobl@tum.de*

6. POSTERS: STATISTICAL GENETICS, GWAS, AND 'OMICS DATA

6a. INVITED POSTER: BAYESIAN FUNCTIONAL GRAPHICAL REGRESSION: APPLICATION TO SMOKING CESSATION STUDIES

Lin Zhang*, University of Minnesota

Veera Baladandayuthapani, University of Texas MD Anderson Cancer Center

Francesco Versace, University of Texas MD Anderson Cancer Center

Jeffrey Morris, University of Texas MD Anderson Cancer Center

We develop a Bayesian functional graphical regression model for multivariate functional data for which functions are collected for each of the p correlated variables within a functional domain. Our method makes inference on graphical Markov models of functional data which allows the graphs to vary over the functional domain. The functional graphical modeling method estimates functional-evolving graphical models in a nonparametric fashion while utilizing a strategy that combines basis function modeling with Bayesian graphical lasso. We show that the functional graphical model that we introduce in the basis space induces a normal scale mixture prior distribution in the data space that leads to shrunken estimators of the precision matrices in the data space. The method borrows strength across functional positions in graphical inference, detects correlations across functional positions, and scales up to large functional datasets collected on a fine grid. We show through simulation and real data analysis that the Bayesian functional graphical regression model can efficiently reconstruct the functional-evolving graphical models by borrowing strength across functional positions.

email: *zhan4800@umn.edu*

6b. INVITED POSTER: TSCAN: PSEUDO-TIME RECONSTRUCTION AND EVALUATION IN SINGLE-CELL RNA-SEQ ANALYSIS

Zhicheng Ji, Johns Hopkins Bloomberg School of Public Health

Hongkai Ji*, Johns Hopkins Bloomberg School of Public Health

When analyzing single-cell RNA-seq data, constructing a pseudo-temporal path to order cells based on the gradual transition of their transcriptomes is a useful way to study gene expression dynamics in a heterogeneous cell population. Currently, a limited number of computational tools are available for this task, and quantitative methods for comparing different tools are lacking. TSCAN is a tool developed to better support in silico pseudo-Time reconstruction in Single-Cell RNA-seq ANalysis. TSCAN uses a cluster-based minimum spanning tree (MST) approach to order cells. Cells are first grouped into clusters and an MST is then constructed to connect cluster centers. Pseudo-time is obtained by projecting each cell onto the tree, and the ordered sequence of cells can be used to study dynamic changes of gene expression along the pseudo-time. Clustering cells before MST construction reduces the complexity of the tree space. This often leads to improved cell ordering. It also allows users to conveniently adjust the ordering based on prior knowledge. TSCAN has a graphical user interface (GUI) to support data visualization and user interaction. Furthermore, quantitative measures are developed to objectively evaluate and compare different pseudo-time reconstruction methods. TSCAN is available at https://github.com/zji90/TSCAN and as a Bioconductor package.

email: *hji@jhu.edu*

6c. EFFICIENT SEMIPARAMETRIC INFERENCE UNDER TWO-PHASE, OUTCOME-DEPENDENT SAMPLING

Ran Tao*, University of North Carolina, Chapel Hill

Donglin Zeng, University of North Carolina, Chapel Hill

Dan-Yu Lin, University of North Carolina, Chapel Hill

The two-phase design is a cost-effective sampling strategy when investigators are interested in evaluating the effects of covariates on an outcome but certain covariates are too expensive to be measured on all study subjects. Under such a design, the outcome of interest and the covariates that are inexpensive to measure are observed for all subjects during the first phase, and the first-

phase information is used to select subjects for measurements of "expensive covariates" during the second phase. In this paper, we consider general two-phase designs, where the outcome of interest can be continuous or discrete, and the "inexpensive covariates" can be continuous and correlated with the expensive covariates. We propose a semiparametric approach to regression analysis by approximating the conditional density functions of expensive covariates given inexpensive covariates with B-spline sieves. We devise a computationally efficient and numerically stable expectation-maximization algorithm to maximize the sieve likelihood. In addition, we establish the consistency, asymptotic normality, and asymptotic efficiency of the resulting estimators. Furthermore, we demonstrate the superiority of the proposed methods over existing ones through extensive simulation studies. Finally, we provide applications to the National Heart, Lung, and Blood Institute Exome Sequencing Project.

email: *dragontaoran@gmail.com*

6d. A STATISTICAL APPROACH TO REMOVE NUISANCE VARIATION IN SINGLE CELL RNA-Seq EXPERIMENTS

Jeea Choi*, University of Wisconsin, Madison
Ning Leng, Morgridge Institute for Research
Li-Fang Chu, Morgridge Institute for Research
Christina Kendziorski, University of Wisconsin, Madison

Oscillatory gene expression is fundamental to mammalian development and aberrations are common in disease. Single-cell RNA sequencing (scRNA-seq) provides the potential to study oscillatory gene expression at an unprecedented scale which provides a major advantage to studies where oscillatory gene expression of interest. However, in many studies, oscillations are not of interest, and the increased variability imposed by them masks the effects that are. To address this, we developed a polynomial regression based method to remove increased variability due to oscillating genes in a snapshot (non time-course) scRNA-seq experiment. Simulation and case studies demonstrate that by removing increased variability due to oscillations, both the power and accuracy of downstream analysis are increased.

email: *jeeachoi@stat.wisc.edu*

6e. CorrMeta: FAST ASSOCIATION ANALYSIS FOR eQTL AND GWAS DATA WITH RELATED SAMPLES AND CORRELATED PHENOTYPES

Kai Xia*, University of North Carolina, Chapel Hill
Andrey A. Shabalin, Virginia Commonwealth University
Wonil Chung, University of North Carolina, Chapel Hill
Zhaoyu Yin, University of North Carolina, Chapel Hill
Martin Styner, University of North Carolina, Chapel Hill
Patrick F. Sullivan, University of North Carolina, Chapel Hill
Fred A. Wright, North Carolina State University
John H. Gilmore, University of North Carolina, Chapel Hill
Rebecca C. Santelli, University of North Carolina, Chapel Hill
Fei Zou, University of North Carolina, Chapel Hill

We develop a computationally efficient alternative, CorrMeta, to a linear mixed-effects model (LMM) for twin genomewide association study (GWAS) data. Instead of analyzing all twin samples together with LMM, CorrMeta first splits twin samples into two independent groups on which multiple linear regression analysis is performed separately, followed by an appropriate meta-analysis to combine the two non-independent test results. Similar idea is also extended to combine GWAS results from multiple correlated phenotypes through CorrMeta. Through mathematical derivations, we prove the validity of CorrMeta. Through simulations, we show empirically that CorrMeta well controlled type I error and negligible power loss compared to the gold linear mixed effects models. Our approaches provide a huge leap in terms of computing power for GWAS data with related subjects and correlated phenotypes. Our method only uses SNP level summary statistics to combine the association analysis of related subjects or correlated phenotypes. Availability: CorrMeta is implemented in R as CorrMeta Twin for twin subjects and CorrMeta MP for multiple phenotypes and are available online.

email: *kxia@email.unc.edu*

6f. NORMALIZATION OF SINGLE CELL RNA-SEQUENCING DATA

Rhonda Bacher*, University of Wisconsin, Madison
Keegan Korthaeur, University of Wisconsin, Madison
Ning Leng, Morgridge Institute for Research
Li-Fang Chu, Morgridge Institute for Research
James A. Thomson, Morgridge Institute for Research
Ron M. Stewart, Morgridge Institute for Research
Christina Kendziorski, University of Wisconsin, Madison

Single cell RNA-sequencing (scRNA-seq) is a promising tool that facilitates study of the transcriptome at the resolution of a single cell. However, along with the many advantages of scRNA-seq come technical artifacts not observed in bulk RNA-seq studies including an abundance of unexpressed genes, varying levels of technical bias across gene groups, and systematic variation in the effects of sequencing depth. The normalization methods traditionally used in bulk RNA-seq were not designed to accommodate these features and, consequently, applying them to the single-cell setting results in poor expression estimates and increased error rates in downstream analyses. To address this, we developed a latent-class non-linear regression framework to enable efficient and accurate scRNA-seq

normalization. Simulation and case study results suggest that the framework provides for improvements in gene expression estimation as well as downstream inference.

email: *rbacher@wisc.edu*

6g. A BAYESIAN HIERARCHICAL MODEL FOR RNA-Seq META-ANALYSIS AND BIOMARKERS CATEGORIZATION BY STUDY HETEROGENEITY

Tianzhou Ma*, University of Pittsburgh

George C. Tseng, University of Pittsburgh

Meta-analysis combining multiple transcriptomic studies increases statistical power and accuracy in detecting differentially expressed genes. As the next-generation sequencing experiments become mature and affordable, increasing number of RNA-seq datasets are available in the public domain. The count-data based technology provides better experimental accuracy, reproducibility and ability to detect low-expressed genes. A naive approach to combine multiple RNA-seq studies is to apply differential analysis tools such as edgeR and DESeq to each study and then combine the summary statistics of p-values or effect sizes by conventional meta-analysis methods. Such a two-stage approach loses statistical power, especially for genes with short length or low expression abundance. In this paper, we propose a full Bayesian hierarchical model (namely, BayesMetaSeq) for RNA-seq meta-analysis by modeling count data, integrating information across genes and across studies, and modeling homogeneous and heterogeneous differential signals across studies. Model-based clustering embedded in the Bayesian model provides categorization of detected biomarkers according to their differential expression patterns across studies that facilitates interpretation and further biological investigation. Simulation and an RNA-seq application on multi-brain-region HIV-1 transgenic rats demonstrate improved sensitivity, accuracy and biological findings of the proposed full Bayesian model.

email: *tim28@pitt.edu*

6h. A NOVEL METHOD FOR TESTING ASSOCIATION WITH COMMON VARIANTS IN CASE-CONTROL STUDIES USING NEXT-GENERATION SEQUENCING DATA

Peizhou Liao*, Emory University

Yijuan Hu, Emory University

Glen A. Satten, Centers for Disease Control and Prevention

Case-control studies with next-generation sequencing (NGS) data are commonly used to detect association between genetic markers and diseases. These studies may suffer from bias due to population stratification. They are also susceptible to confounding effects of systematic differences in read depths and sequencing errors between cases and controls, especially when an external NGS control group is used. To assess the association in sequencing data, we propose a novel method to test association using read count data only. We first develop an unbiased estimator of the true genotype using read data, and then test for differences between cases and controls using a robust variance estimator. In addition, we present a method to estimate principal components from read count data which can be used to correct for population stratification. Our approach is computationally simple and generally applicable to whole-genome association studies of common variants. We perform extensive simulations to demonstrate that our method is effective in controlling type I error and has comparable power to a more computationally-intensive likelihood-based approach. We apply our approach to a NGS data set to evaluate its performance in real applications.

email: *pliao3@emory.edu*

6i. NOVEL TESTS FOR DETECTION OF GENE-ENVIRONMENT INTERACTION IN FAMILY STUDIES

Brandon J. Coombes*, University of Minnesota

Saonli Basu, University of Minnesota

In this paper, we consider testing for interactions between a group of genetic variants and a set of environmental factors within a family study. While testing these interactions one-by-one in separate models may be straightforward, we may lose power to detect association of an interaction with disease. In this paper, we extend recently developed score-based methods for genetic associations to the gene-environment interaction testing problem. We also extend a variance component score test of the interactions to family studies. We additionally propose novel methods based on a sequential scoring algorithm and compare the performance among all of the methods. Finally, a Lasso regression pre-step is proposed under the null hypothesis of no gene-environment interaction to filter out null genetic variants and thus reduce the number of likely null interactions. Our simulations show that this filtering step produces large power gains, especially with our proposed main-effect scoring methods. We also see that our scoring methods do very well if the interactions are in the same direction, even in situations where there is sparse interaction.

email: *coom0054@umn.edu*

6j. INCORPORATING BIOLOGICAL INFORMATION IN SPARSE PRINCIPAL COMPONENT ANALYSIS WITH APPLICATION TO GENOMIC DATA

Ziyi Li*, Emory University

Sandra Safo, Emory University

Qi Long, Emory University

The advances in technology have lead to the collection of high dimensional data such as genomic data. Before applying the existing statistical methods on high dimensional data, principal component analysis (PCA) is often used to reduce dimensionality. Sparse PC loadings are usually desired in this situation for simplicity and better interpretation. Although PCA has been extended to produce sparse PC loadings, few methods take potential biological information into consideration. In this article, we propose two novel structured sparse PCA methods which not only have sparse solutions but also incorporate available biological information. Our simulation study demonstrates incorporating known biological information improves the performance of sparse PCA methods, and the proposed methods are robust to potential misspecification of the biological information. We further illustrate the performance of our methods in a Glioblastoma genomic data set.

email: *Ziyi.li@emory.edu*

6k. NOVEL THEORY FOR MAPPING AND CHARTING THE GENETIC ARCHITECTURE OF GENE EXPRESSION PROFILES ON MULTIPLE TISSUES

Kirk Gosik*, Penn State College of Medicine

Rongling Wu, Penn State College of Medicine

A novel theory for mapping and charting the genetic architecture of gene expression profiles on multiple tissues and further develop an algorithmic platform for identifying expression quantitative trait loci (eQTLs) that regulate coordinated expression of genes from different tissues is proposed. The central idea of our theory is to view a living organism as an evolutionary system in which different tissues interact and coordinate with each other through the compromised expression of genes driven by two opposite mechanisms, competition and cooperation. We pursue to develop statistical approaches for mapping eQTLs expressed over multiple tissues and identifying those that participate in the coordination of gene expression from different tissues. First, we integrate the widely used game-theoretic model into association studies to map eQTLs that control complex interactions between tissues and explain altruistic behaviors in terms of Darwinian competition. We incorporate control theory into eQTL mapping by a system of differential equations, which enables the quantitative characterization and prediction of the dynamic changes of genetic effects on different tissues. Third, we employ and reform functional mapping, a dynamic framework for QTL mapping, to characterize how eQTLs are expressed differently over different tissues and chart an overall picture of eQTL tissue interactions.

email: *kgosik@hmc.psu.edu*

6l. ESTIMATING CELL TYPE SPECIFIC ASSOCIATIONS FROM WHOLE BLOOD METHYLATION

Richard T. Barfield*, Harvard University

Xihong Lin, Harvard University

Association analysis of DNA methylation (DNAm) data is challenged by cell type heterogeneity, as the data is typically a mixture of cell types. Cell type heterogeneity can bias results since DNAm is a mechanism in tissue and cell differentiation. To correct for this, analyses include observed or estimated cell type counts as covariates. This does not however estimate exposure effects on cell type specific DNA methylations. Direct measurements of cell type specific methylation would involve costly lab work. We develop here a statistical method to estimate cell specific associations using whole blood methylation data when cell composition is available but cell-specific methylations are not. We assume cell type specific regression models of the exposure effects on cell type specific methylations. We treat cell specific methylations as missing, and develop an EM algorithm to estimate the cell specific exposure effects using whole blood methylation data and cell type counts. We analyzed data from the Normative Aging Study to examine cell specific smoking associations on 49 probes established to be associated with smoking. Five probes had a statistically significant cell type specific association. To the best of our knowledge, this is the first method to estimate these effects from whole blood.

email: *rbarfield01@fas.harvard.edu*

6m. APPLICATION OF SAMPLE QUALITY WEIGHTS IN RANDOM EFFECTS META-ANALYSIS OF GENE EXPRESSION STUDIES: BAYESIAN AND NON-BAYESIAN APPROACHES

Uma Siangphoe*, Virginia Commonwealth University

Nitai D. Mukhopadhyay, Virginia Commonwealth University

Study heterogeneity in meta-analysis of gene expression studies is usually unknown and can arise from inconsistency of experimental conditions and sample quality issues. Random effects meta-analysis models are commonly applied to handle the study heterogeneity. High heterogeneity tends to be detected in microarray studies that contain low quality samples, which can reduce statistical power of the models. Due to the sample quality issues, we developed a meta-analytic approach that includes sample quality weights to adjust the study heterogeneity in the random effects meta-analysis models. Twenty-one sample quality weights were implemented in the standard random effects models, some of the appropriate weights were selected to implement through the Bayesian approach. Bayesian hierarchical models with different prior distributions for study heterogeneity were examined. The performance of random effects meta-analysis models with and without sample quality weights in

the classical and Bayesian approaches were compared using a series of simulation studies. We demonstrate that an optimal random effects meta-analysis model in the classical approach performs similarly to an optimal model in the Bayesian approach. Most of the sample quality weights increase precision of the random effects meta-analysis models as compared to the non-weighted models.

email: *siangphoeu@vcu.edu*

6n. INTERMITTENCY AND LIMIT THEOREMS FOR SUPERPOSITIONS OF ORNSTEIN-UHLENBECK TYPE PROCESSES

Danijel Grahovac, University of Osijek

Nikolai Leonenko, Cardiff University

Alla Sikorskii, Michigan State University

Irena Tesnjak*, Michigan State University

Stationary Ornstein-Uhlenbeck (OU) type processes driven by Levy noise have been extensively used in modeling of high frequency (genetic and financial) data. Discrete superpositions of these processes can be constructed to incorporate non-Gaussian marginal distributions and long or short range dependence. While the partial sums of finite superpositions of OU type processes obey the central limit theorem, the partial sums of infinite long range dependent superpositions are intermittent. We discuss the property of intermittency and show that intermittent behavior precludes central limit theorem type results for the partial sums.

email: *tesnjaki@st.msu.edu*

6o. POPULATION GENETIC FEATURES OF RARE VARIANTS IN FINLAND

Rosemary Putler*, University of Michigan

Sebastian Zoellner, University of Michigan

The population of Finland has an interesting, and well-documented, demographic history: multiple small waves of migration into Finland followed by a rapid population expansion. This history leaves a distinct pattern in Finnish genomes including an excess of rare heritable diseases and a slightly reduced genetic heterogeneity among common variants. However, the impact of this population history on rare variants is understood less well. It is not clear how much rare variation Finns share with other European countries or how much rare variant diversity is affected by population structure. However, these questions are critical for effectively leveraging samples from Finland in sequencing studies. Here we examine a sample of 2132 exomes from Finnish individuals, and quantify the differentiation of rare variation between Finns and other European samples. We examined allele sharing and variance component statistics between these populations and show that within rare variants, Finns are more similar to each other than to Swedish and British populations. We further illustrate substantial population structure in Finland and create a reference map for the geographic distribution of Finnish genotypes that shows correlation between genetic variation and geographic origin. This map can be used to control population stratification in rare variant tests in Finland.

email: *rputler@umich.edu*

6p. METHODS OF INFERENCE FOR PENALIZED REGRESSION IN HIGH-DIMENSIONAL GENETIC ASSOCIATION STUDIES

Jaron Arbet*, University of Minnesota

Saonli Basu, University of Minnesota

In high-dimensional "p>n" settings where the number of predictors exceeds the sample size, multiple linear regression with Ordinary Least Squares estimation is no longer viable. Penalized regression is an attractive alternative to estimate regression coefficients in such settings. Today, GWAS and other high-dimensional genetic association studies are predominantly analyzed using Single Marker Association methods (SMA) - where one analyzes the marginal relationship between a single predictor and the response of interest. Penalized regression allows one to simultaneously model the relationship between thousands of genetic predictors and the response with a single model. Thus penalized regression more realistically models the underlying genetic architecture, and may increase power due to decreased residual variance or the presence of interaction among causal predictors. Several penalty types are discussed and methods for tuning the penalty parameters; then various methods for conducting inference on the penalized coefficients that allow for Type-1-Error or False Discovery Rate control are discussed and compared via simulations. Overall, whether one wishes to control the Type-1-Error or FDR, penalized regression is often significantly more powerful than SMA. In particular, using permutations to select tuning parameters that control Type-1-Error is consistently the most powerful method, and is a computationally efficient alternative to SMA.

email: *arbet003@umn.edu*

6q. A POWERFUL APPROACH IN DIFFERENTIAL ANALYSIS FOR TIME SERIES MICROBIAL STUDIES

Dan Luo*, University of Arizona

Lingling An, University of Arizona

Metagenomics has a great potential to discover previously unattainable information about microbial communities. Detecting differentially abundant features (e.g., species or genes) plays a critical role in revealing the contributors (i.e., pathogens) to the status (e.g., disease) of microbial samples. However, currently available statistical methods lack power in detecting differentially abundant

features across different conditions, in particular, for time series metagenomic data. We have proposed a novel procedure to meet with the challenges in detecting differentially abundant features from metagenomic samples under different biological/medical conditions. The new approach takes advantage of dependence structure of time series data and the detection procedure relies on sound statistical support. Not only it can accurately identify the different features but also result in the information on the start and end time points. Compared with other existing methods the new approach shows the best performance in the comprehensive simulation studies. The new method is also applied to real metagenomic datasets and the new interesting findings may provide another angle of understanding the mechanism of the diseases.

email: *luodan@email.arizona.edu*

6r. A TWO-PART MIXED EFFECT MODEL FOR LONGITUDIAL MICROBIOME DATA ANALYSIS

Eric Z. Chen*, University of Pennsylvania

Hongzhe Li, University of Pennsylvania

The human microbial communities are associated with many human diseases such as obesity, diabetes and inflammatory bowel disease. High-throughput sequencing technology has been widely used to quantify the microbial composition in order to understand their impact on human health. The longitudinal design is quite common in many microbiome studies. A key question in the microbiome study is to identify the microbes that are associated with clinical outcomes or environmental factors. However, the longitudinal microbiome compositional data are highly skewed, bounded in [0,1], and often sparse with many zeros. Moreover, the data from repeated measures are correlated. Therefore, a method that takes into account these features is needed for association analysis of longitudinal microbiome data. In this paper, we propose a two-part mixed effect model to identify association between microbial abundnace and clinical covariates for longitudial microbiome data. The model includes a logistic component to model present/absent of the microbe and a beta component to model non-zero microbial abundance. Each component involves a random effect to take into account the correlation among repeated mearsuments on the same subject. Simulation studies show that the ZIBR model outperforms the commonly used method. We also applied the ZIBR model to the real data from human gut microbiome study and the results show good consistent with the literactures. We provide a good tool for association analysis in microbiome research.

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

6s. A TWO-STEP INTEGRATED APPROACH TO DETECT DIFFERENTIALLY EXPRESSED GENES IN RNA-Seq DATA

Naim A. Mahi*, University of Cincinnati

Munni Begum, Ball State University

Ribonucleic acid sequencing or RNA-Seq experiments produce millions of discrete DNA sequence reads, as a measure of gene expression levels. It enable researchers to investigate complex aspects of the genomic studies. One of the common assumptions of RNA-Seq data is that, all gene counts follow an overdispersed Poisson or negative binomial distribution which is sometimes misleading because within each treatment group, some genes may have constant transcription levels with no overdispersion. In such cases, it is more appropriate to consider two sets of genes: overdispersed and non-overdispersed. We propose a new two-step integrated approach to detect differentially expressed genes in RNA-Seq data using standard Poisson model for non-overdispersed genes and NB model for overdispersed genes. This is an integrated approach because this method can be combined with any other NB based methods for detecting DE genes. We evaluate the proposed approach using two simulation strategies and two real RNA-Seq data. We compare the performance of our proposed method combined with the three popular R-software packages namely edgeR, DESeq, and NBPSeq with their default settings. For both the simulated and real data sets, integrated approaches perform better or at least equally well compared to the regular methods embedded in these R-packages.

email: *mahina@mail.uc.edu*

7. POSTERS: METHODOLOGY AND APPLICATIONS IN EPIDEMIOLOGY, ENVIRONMENT, AND ECOLOGY

7a. INVITED POSTER: FALSE DISCOVERY RATE SMOOTHING

James G. Scott*, University of Texas, Austin

Wesley Tansey, University of Texas, Austin

We present false discovery rate smoothing, an empirical-Bayes method for exploiting spatial structure in large multiple-testing problems. FDR smoothing automatically finds spatially localized regions of significant test statistics. It then relaxes the threshold of statistical significance within these regions, and tightens it elsewhere, in a manner that controls the overall false-discovery rate at a given level. This results in increased power and cleaner spatial separation of signals from noise. The approach requires solving a non-standard high-dimensional optimization problem, for which an efficient augmented-Lagrangian algorithm is presented. We demonstrate that FDR smoothing exhibits state-of-the-art performance on simulated

examples. We also apply the method to a data set from an fMRI experiment on spatial working memory, where it detects patterns that are much more biologically plausible than those detected by existing FDR-controlling methods. All code for FDR smoothing is publicly available in Python and R.

email: *james.scott@mccombs.utexas.edu*

7b. ASYMPTOTIC BEHAVIORS OF THE MANTEL-HAENSZEL ESTIMATORS AND THEIR ROBUST VARIANCE ESTIMATORS WHEN THE COMMON EFFECT ASSUMPTIONS ARE VIOLATED

Hisashi Noma*, The Institute of Statistical Mathematics

Kengo Nagashima, Chiba University

The Mantel-Haenszel estimators for the common effect parameters of stratified 2×2 tables have been widely adopted in epidemiological and clinical studies. Although the Mantel-Haenszel estimators are simple and effective estimating methods, the correctness of the common effect assumptions cannot be justified in general practices. And then, the targeted "common effect parameters" do not exist. Under these settings, even if the Mantel-Haenszel estimators have desirable properties, it is quite uncertain what they estimate and how the estimates are interpreted. In this study, we conducted theoretical evaluations for their asymptotic behaviors when the common effect assumptions are violated. We explicitly showed that the Mantel-Haenszel estimators converge to weighted averages of stratum-specific effect parameters and they can be interpreted as intuitive summaries of the stratum-specific effect measures. Also, the Mantel-Haenszel estimators correspond to some standardized effect measures under certain conditions. In addition, we developed robust variance estimators which are valid even when the common effect assumptions are violated. We implemented numerical studies based on several epidemiologic studies for evaluating empirical properties of these estimators, and confirmed general validities of these theoretical results.

email: *noma@ism.ac.jp*

7c. StatStart, HARVARD UNIVERSITY BIOSTATISTICS DEPARTMENT

Octavious Talbot*, Harvard University

Sam Tracy, Harvard University

Alex Ocampo, Harvard University

Underrepresented minorities, which include Blacks, Hispanics, and American Indian/Alaskan Natives, comprised 36.3% of the US population aged 18-24 in 2012 [1]. However, this demographic only accounted for 11.5% of all STEM doctoral degrees received in 2012 and only 6.2% of all college faculty positions in 2013 [2]. In an effort to address this disparity, we developed StatStart, a statistical programming program to inspire Boston area underrepresented minority high school students to pursue a career in STEM. In the summer of 2015, the inaugural summer of StatStart, ten students traveled to the Harvard Biostatistics department every day for a month to learn introductory statistics and programming in R. In addition to course material, small groups of students deeply explored a specific statistical topic alongside a graduate student mentor. Topics included modeling herd immunity, the central limit theorem, weak law of large numbers, as well as analyzing public health datasets from the World Bank and CDC. The program culminated in a formal presentation to an audience of Biostatistics students, faculty, and StatStart participant parents. We hope that sharing our experiences and exchanging ideas with like-minded colleagues at ENAR will inspire other institutions to implement similar programming.

email: *octavioustalbot@g.harvard.edu*

7d. A BAYESIAN APPROACH TO ACCOUNT FOR MISCLASSIFICATION AND OVERDISPERSION IN COUNT DATA

Wenqi Wu*, Baylor University

James Stamey, Baylor University

David Kahle, Baylor University

Count data are subject to considerable sources of what is often referred to as non-sampling error. Errors such as misclassification, measurement error and unmeasured confounding can lead to substantially biased estimators. It is strongly recommended that epidemiologists not only acknowledge these sorts of errors in data, but incorporate sensitivity analyses into part of the total data analysis. We extend previous work on Poisson regression models that allow for misclassification by thoroughly discussing the basis for the models and allowing for extra-Poisson variability in the form of random effects. Via simulation we show the improvements in inference that are brought about by accounting for both the misclassification and the overdispersion.

email: *wenqi_wu@baylor.edu*

7e. IMPROVING THE DYNAMICS OF DATA-DRIVEN DISCOVERY AT ACADEMIC HEALTH CENTERS

Jonathan Gelfond, University of Texas Health Science Center, San Antonio

Martin W. Goros*, University of Texas Health Science Center, San Antonio

Problem: Translational research conducted at academic health centers (AHC) has become increasingly data intensive. In order to effectively and efficiently operate in this research domain, biostatistics, and research design resource units are required to provide expertise and training for designing, conducting and ana-

lyzing studies. However, this demand for services strains existing resources, which may lead to decreases in productivity as well as increased costs. Approach: We applied lean six sigma principles to establish a systematic continual process improvement cycle for translational research data analysis. The goal of which is to assess and improve the efficiency and effectiveness of the BERD unit operations by objectively measuring outcomes and the impact of interventions in policies and processes. We used the define, measure, analyze, improve, and control (DMAIC) methodology. The required steps of this method are to define a measures process of data analysis in this context, measuring these using a web-based system, analyzing these operational metrics, implementing improvements, and controlling the process. Outcomes: We defined operational states for the data analysis process. These states are Intake, Data Validation, Analysis, and Reporting. These measures were collected over a 6 month period, and a web-based dashboard application was built to record and analyze these states. This dashboard established an operational awareness that led to innovations in processes. These innovations led to measureable improvements in operational metrics.

email: *goros@uthscsa.edu*

7f. SIMULTANEOUS PREDICTION OF ANTICANCER ACTIVITY AND TOXICITY IN ALIPHATIC NITROSOUREAS USING QUANTITATIVE STRUCTURE ACTIVITY RELATIONSHIP (QSAR) METHODS

Soyi Sarkar*, Newark Academy

Isabel Alland, Newark Academy

Solomon H. Jacobson, Newark Academy

Developing anticancer drugs with low toxicity and high therapeutic activity can be challenging. Predictive statistical models, such as QSAR, are useful for identifying new target molecules with a desired set of properties. We present a new QSAR study of aliphatic nitrosourea molecules which have been used to treat hematologic malignancies. QSAR models were developed to predict both anticancer activity (ACA) and toxicity values of nitrosourea molecules. Our QSAR models used both 2-D and 3-D descriptors calculated with eDragon. JMP was used for objective feature selection and neural net modeling. The resulting models had high correlation coefficients. The model R2 for actual vs. predicted ACA was 0.98, and the cross validated R2 was 0.96. The R2 values for the acute toxicity model were 0.96 and 0.93, respectively. Model predictions were then used for identification of nitrosoureas which may be safe and effective against tumors. Future studies can focus on screening other new untested nitrosoureas.

email: *ssarkar17@newarka.edu*

7g. INCORPORATING CANDIDATES WITH MULTIPLE ASSOCIATED INCOMPATIBLE DONORS IN KIDNEY PAIRED-DONATION

Mathieu Bray*, University of Michigan

Wen Wang, University of Michigan

Peter X-K. Song, University of Michigan

John D. Kalbfleisch, University of Michigan

Kidney paired-donation (KPD) represents one possible avenue for transplant candidates seeking a donor. In KPD, a candidate with a willing incompatible donor is matched with other such pairs in an effort to find combinations of donor exchanges that allow all candidates involved to obtain transplants. Finding successful exchange combinations within KPD, however, is often limited in practice. In some cases, candidates have several donors, all incompatible but willing to participate in KPD. Having multiple associated donors in KPD introduces additional opportunities to match with other pairs, and also allows for possibilities to fall back to immediate alternatives should any failure (eg. withdrawal, laboratory test overturning the presumed compatibility of a match) occur in a determined transplant arrangement. Exchanges involving pairs with multiple associated donors should be preferred in selection within the network of possible KPD solutions. We formulate an objective assignment of expected utility for KPD exchange combinations, both with and without recourse to available fallback options and accounting for probabilities of failure, where candidates can have multiple associated incompatible donors. This extends previous mathematical formulations of the KPD problem (Li, 2012). Further, we illustrate through simulation the benefits for candidates in seeking out multiple donors for KPD.

email: *braymath@umich.edu*

7h. ESTIMATING NEIGHBORHOOD SOCIOECONOMIC STATUS INDEXES IN CANCER RISK MODELS

David C. Wheeler*, Virginia Commonwealth University

Jenna Czarnota, Virginia Commonwealth University

Mary H. Ward, National Cancer Institute, National Institutes of Health

Socioeconomic status (SES) is often considered as a risk factor for disease. SES is typically measured using a combination of educational attainment, income, and employment variables and represented as a composite variable. Approaches to building the composite variable include using arbitrary weights for each variable or estimating the weights with principal components analysis (PCA) or regression analysis. However, PCA does not consider the relationship between the health outcome and the SES variables when constructing the index. Standard regression methods may suffer from collinearity effects when estimating the variable weights

due to strong correlation between the SES variables. In this project, we use weighted quantile sum (WQS) regression to estimate a neighborhood level SES index in a model of risk of non-Hodgkin lymphoma (NHL) in the NCI-SEER NHL study. We consider historic neighborhood SES variables from different decades of the US Census by linking residential histories collected in the study to census block group to account for disease latency. An advantage of WQS regression in this setting is that it can estimate the effect of an SES index and the weights for each variable in the index while limiting effects of collinearity.

email: *dcwheeler@vcu.edu*

7i. COMPARISON OF LINEAR, QUADRATIC, AND LINEAR SPLINE REGRESSION MODELS TO EXAMINE THE RELATIONSHIP BETWEEN BIRTH-WEIGHT AND SYSTOLIC BLOOD PRESSURE IN CHILDREN

Amna Umer, West Virginia University
Candice Hamilton, West Virginia University
Cris Britton, West Virginia University
Lee Pyles, West Virginia University
William Neal, West Virginia University
Collin John, West Virginia University
Christa Lilly*, West Virginia University

Few studies have suggested a 'U' shape relationship between birth weight (BTW) and later systolic blood pressure (SBP); however, many studies continue to utilize linear relationships to examine this association. The objective of the study was to examine this relationship with linear, quadratic, and linear spline regression models. The study used longitudinally linked data from two cross-sectional surveillance datasets in rural Appalachian children (N=22,136). A simple linear regression was performed between BTW and SBP, and a quadratic term added. Linear spline regression analysis was performed by adding first a knot at BTW <2500g, and then a second knot at BTW>4000g, conventional cut-off criteria for BTW. Additional knots were also explored. The study compared the mean squared error (MSE) and R2values. The MSE (12.05) was lowest and R2highest (0.0008) for the spline regression model that included knots at traditional low and high BTW cut-offs. The results of the quadratic regression model were similar to the spline model. The linear regression model had the highest MSE (12.054) and lowest R2value (0.0002). We demonstrate that the spline model with two knots (low and high cut-offs) is the most appropriate model to use when examining this relationship in epidemiological studies.

email: *cice@hsc.wvu.edu*

7j. AN ACTIVITY INDEX FOR RAW ACCELEROMETRY DATA AND ITS COMPARISON WITH ACTIVITY COUNTS

Jiawei Bai*, Johns Hopkins University
Chongzhi Di, Fred Hutchinson Cancer Research Center
Luo Xiao, North Carolina State University
Kelly R. Evenson, University of North Carolina, Chapel Hill
Andrea Z. LaCroix, University of California, San Diego
Ciprian M. Crainiceanu, Johns Hopkins University
David M. Buchner, University of Illinois, Urbana-Champaign

Technology advances have allowed densely sampled raw accelerometry data to be collected and stored in many recent studies. However, limited effort has been made to extract useful information from such data. In this work we 1) proposed the physical activity index (AI), a new metric for summarizing raw tri-axial accelerometry data and 2) compared the AI to the activity count's (AC) predictive performance for estimating energy expenditure and various types of activities. The AI was defined to be the variability of raw acceleration signals, after being normalized by the systematic noise of the accelerometer. The AI has 3 important properties: ease to deploy, additivity and invariance to rotation. The AI was compared with AC generated by accelerometer GTX3+, for distinguishing among various types of activities and predicting energy expenditure via receiver operating characteristic curve analyses and area under the curve. The result revealed that AI had a much improved prediction performance both for distinguishing various types of activities and for predicting energy expenditure. The AI can not only be used as a replacement of AC within current analysis framework, but also facilitate comparing results of studies using different brands of accelerometer.

email: *javybai@gmail.com*

7k. SECONDARY RESPONSE VARIABLE REGRESSION ANALYSIS IN A CASE-COHORT STUDY

Yinghao Pan*, University of North Carolina, Chapel Hill
Haibo Zhou, University of North Carolina, Chapel Hill
Jianwen Cai, University of North Carolina, Chapel Hill
Sangmi Kim, Georgia Regents University

Case-cohort design has been widely used to reduce the cost for a time-to-event study. In any real study, there are typically more than one endpoints. Researchers often would like to reuse the available case-cohort data to study the relationship of a secondary endpoint with the primary exposure obtained in the case-cohort study. Because the case-cohort sample is not a random sample from the general population, how to perform the secondary outcome analysis correctly and efficiently is a challenging, yet must faced hurdle for many investigators. In this paper, we proposed an estimated

likelihood approach for analyzing the secondary outcome in a case-cohort study. The estimation is based on maximizing a semiparametric likelihood function that is built jointly on both time-to-failure outcome and the secondary outcome. The proposed estimator is shown to be consistent, efficient and asymptotically normal. Finite sample performance is evaluated via simulation studies, and empirical data from the Sister Study is presented to illustrate our method.

email: *yypan@live.unc.edu*

7I. THE ALERT ALGORITHM FOR DETECTION OF LOCAL ONSETS OF RSV AND INFLUENZA

Alexandria C. Brown*, University of Massachusetts, Amherst

Nicholas G. Reich, University of Massachusetts, Amherst

The Centers for Disease Control (CDC) annual estimates of influenza-associated deaths range from 3,349 in 1986-1987 to 48,614 in 2003-2004 in the United States. A report on the mortality caused by these diseases showed that influenza A (H3N2) caused the highest number of deaths, followed by RSV, influenza B, and influenza A (H1N1) viruses. Many of these deaths were young, elderly, or immunocompromised individuals, indicating that outbreak detection at the community-level healthcare setting may be necessary to reduce the mortality of these respiratory viruses. One method for identifying a local shift toward epidemic activity is the Above Local Elevated Respiratory Illness Threshold (ALERT) algorithm. We will validate ALERT for use in 1) multi-strain datasets, such as influenza A and B combined and 2) non-influenza respiratory viruses, such as RSV. In some cases, variable findings were attributable to differences in seasonality, noise, and trend characteristics of each dataset. We test and confirm these findings using a simulation study. Finally, we evaluate and discuss our results to provide recommendations for dataset attributes when applying ALERT to local case count data.

email: *acbro0@schoolph.umass.edu*

8. POSTERS: VARIABLE SELECTION AND METHODS FOR HIGH DIMENSIONAL DATA

8a. INVITED POSTER: FLEXIBLE MODELING AND FEATURE IMPORTANCE IN HIGH DIMENSIONAL PROBLEMS

Noah Simon*, University of Washington

In this poster we will discuss methods for flexible modeling with structure in high dimensional problems. In particular we will focus on penalized regression. We will discuss how to combine penalties to induce multiple structures of interest, and what effect this has on computation and asymptotic convergence. In addition we will propose a parameter which reflects the importance of a feature; and give some asymptotic results for estimating this parameter.

email: *nrsimon@uw.edu*

8b. SINGLE-INDEX VARYING COEFFICIENT MODEL FOR FUNCTIONAL RESPONSES

Xinchao Luo*, East China Normal University and University of North Carolina, Chapel Hill

Lixing Zhu, Hong Kong Baptist University

Hongtu Zhu, University of North Carolina, Chapel Hill

The aim of this paper is to develop a single-index varying coefficient (SIVC) model for establishing a varying association between functional responses (e.g., image) and a set of covariates. It enjoys several unique features of both varying-coefficient and single-index models. A procedure is developed to estimate varying coefficient functions, link function, and the covariance function of individual functions. The optimal integration of information across different grid points are systematically delineated and the asymptotic properties (e.g., consistency and convergence rate) of all estimators are examined. Simulation studies are conducted to assess the finite-sample performance of the proposed procedure. Furthermore, our real data analysis of a white matter tract dataset obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) study confirms the advantage and accuracy of SIVC model over the popular varying coefficient model.

email: *fallenstar0909@gmail.com*

8c. ASSESSMENT OF DPOAE TEST-RETEST DIFFERENCE CURVES VIA HIERARCHICAL GAUSSIAN PROCESSES

Junshu Bao*, University of South Carolina

Timothy E. Hanson, University of South Carolina

Distortion product otoacoustic emissions (DPOAE) testing is a promising alternative to behavioral hearing tests and auditory brainstem response testing of pediatric cancer patients. The central goal of this study is to assess whether significant changes in the DPOAE frequency/emissions curve (DP-gram) occur in pediatric patients in a test-retest scenario. This is accomplished through the construction of normal reference charts, or credible regions, that DP-gram differences lie in, as well as contour probabilities that measure how abnormal (or in a certain sense rare) a test-retest difference is. A challenge is that the data were collected over varying frequencies, at different time points from baseline, and on possibly one or both ears. A hierarchical structural equation Gaussian process model is proposed to handle the three sources of correlation in the emissions measurements, wherein both subject-specific random effects and variance components governing the smoothness and variability of each child's Gaussian process are coupled together.

email: *bao3@email.sc.edu*

8d. VARIABLE SELECTION IN FUNCTION-ON-SCALAR REGRESSION

Yakuan Chen*, Columbia University

Jeff Goldsmith, Columbia University

Todd Ogden, 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 aB-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 establish theoretical properties for the resulting estimator. We further 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: *yc2641@cumc.columbia.edu*

8e. VARIABLE SCREENING IN MULTICATEGORY CLASSIFICATION

Yue Zeng*, University of Arizona

Hao Helen Zhang, University of Arizona

Ning Hao, University of Arizona

Classification with high-dimensional features are commonly encountered in many scientific problems in biology, genetics, medicine, and so on. When the number of features is ultra high, a fast and effective dimension reduction is needed or desired to capture important signals, filter out noises, and down-scale the data set without information loss, before a refined and more computationally expensive analysis. In this paper, we study the problem variable screening in multicategory classification problems. A variety of screening procedures are considered, including likelihood-based and LDA procedures, along with screening methods based on pairwise classification. These tools are thoroughly evaluated and compared at various scenarios, and then applied to cancer classification.

email: *zengyue@email.arizona.edu*

8f. INFER EDGE STRUCTURE OF MIXED GRAPHIC MODEL

Suwa Xu*, University of Florida

Gaussian Graphic Models are widely used to explore networks, such as gene regulatory networks. However typical data source contain both continuous and discrete data. We present a new measure of correlation structure on mixed data. Under some mild assumptions, our reduced-size correlation model is equivalent to the true correlation model in construction of mixed graphic models. Further, we establish the consistency of the proposed method. Simulations as well as real data examples are given to illustrate the efficiency and flexibility of our method.

email: *suwaxu@ufl.edu*

8g. BI-LEVEL VARIABLE SELECTION IN AN ORDERED PROBIT REGRESSION MODEL VIA MAXIMUM LIKELIHOOD WITH COMPOSITE BRIDGE PENALTY

Feiran Jiao*, University of Iowa

Kung-sik Chan, University of Iowa

We study the problem of variable selection in an ordered probit model with an ordinal response variable. In practice, the covariates may often be grouped but some groups may be mixed in that they contain both relevant and irrelevant variables, i.e. whose coefficients are non-zero and zero, respectively. Thus, it is pertinent to develop a consistent method for simultaneously selecting relevant groups and the relevant variables within each selected group, which constitutes the so-called bi-level selection problem. We propose to use a penalized maximum likelihood approach with a composite bridge penalty to solve the bi-level selection problem in an ordered probit model. An EM algorithm is developed for implementing the proposed method. The proposed approach is shown to enjoy a number of desirable theoretical properties including bi-level selection consistency and oracle properties, in both classical and high-dimensional settings. Simulations demonstrate that the proposed method enjoys good empirical performance. The method is illustrated with a real medical application. The ordered probit model assumes that the latent response is conditionally normal; the proposed method can be extended to the case of conditionally non-normal latent response.

email: *feiran-jiao@uiowa.edu*

8h. FSEM: FUNCTIONAL STRUCTURAL EQUATION MODEL FOR TWIN FUNCTIONAL DATA

Shikai Luo*, North Carolina State University

Rui Song, North Carolina State University

Martin Styner, University of North Carolina, Chapel Hill

John Gilmore, University of North Carolina, Chapel Hill

Hongtu Zhu, University of North Carolina, Chapel Hill

The aim of this paper is to develop a novel class of functional structural equation models (FSEMs) to dissect functional genetic and environmental effects on twin functional data, while characterizing the varying association between functional data and covariates of interest. We propose a three-stage efficient estimation procedure to estimate varying coefficient functions for various covariates (e.g., gender) as well as two covariance operators for the genetic and environmental effects. We develop an inference procedure based on weighted likelihood ratio statistics to test the genetic/environmental effect at either a fixed location or a compact region. We also systematically carry out the theoretical analysis of the estimated varying functions, the weighted likelihood ratio statistics, and the estimated covariance operators. We conduct extensive Monte Carlo simulations to examine the finite-sample performance of the estimation and inference procedures. We apply the proposed FSEM to model the genetic and environmental effects on twin white-matter tracts obtained from the UNC early brain development study.

email: *sluo@ncsu.edu*

8i. SIFORM: SHARED INFORMATIVE FACTOR MODELS FOR INTEGRATION OF MULTI-PLATFORM BIOINFORMATIC DATA

Xuebei An*, 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

High-dimensional omic data derived from different technological platforms have been extensively used to facilitate comprehensive understanding of disease mechanisms. Numerous studies have integrated multi-platform omic data; however, few have efficiently and simultaneously addressed the problems that arise from high dimensionality and complex correlations. We propose a statistical framework of shared informative factor models that can jointly analyze multi-platform omic data and explore their associations with a disease phenotype. The common disease-associated sample characteristics across different data types can be captured through the shared structure space, while the corresponding weights of genetic variables directly index the strengths of their association with the phenotype. Extensive simulation studies demonstrate the performance of the proposed method in terms of biomarker detection accuracy via comparisons with three popular regularized regression methods. We also apply the proposed method to The Cancer Genome Atlas lung adenocarcinoma data set to jointly explore associations of mRNA expression and protein expression with smoking status. Many of the identified biomarkers belong to key pathways for lung tumorigenesis, some of which are known to express differently across smoking levels. We discover potential biomarkers that reveal different mechanisms of lung tumorigenesis between light smokers and heavy smokers.

email: *xuebei.an@gmail.com*

8j. HYPOTHESIS TESTING FOR TIME-VARYING COVARIATE EFFECT IN COMPLEX CORRELATED FUNCTIONAL DATA

Saebitna Oh*, North Carolina State University

Ana-Maria Staicu, North Carolina State University

We consider complex correlated functional data where are observed at multiple instances (often visit times) per subject, for many subjects. Our interest is to develop inferential methods to study the population effect of covariates in this setting. We propose a pseudo F- testing procedure that accounts for the complex error structure and is computationally efficient. We use a transformation through functional principal component analysis and eigen decomposition of error covariance. Simulation studies confirm that the testing approach has the correct size and compares favorably with available competitors in terms of power. The methods are evaluated on a data application.

email: *soh3@ncsu.edu*

8k. ThrEEboost: THRESHOLDED BOOSTING FOR VARIABLE SELECTION AND PREDICTION VIA ESTIMATING EQUATIONS

Benjamin T. Brown*, University of Minnesota

Christopher J. Miller, 3D Communications

Julian Wolfson, University of Minnesota

Most variable selection techniques for high-dimensional models are designed to be used in settings where observations are independent, completely observed, and otherwise “clean”. There are other approaches to estimate low-dimensional parameters in the presence of correlation, measurement error, and otherwise “messy” data. We present ThrEEBoost (ThresholdedEstimatingEquationBoosting), an iterative, boosting-like technique which allows high-dimensional variable selection to be performed by solving an estimating equation. A thresholding parameter controls the number of coefficient values updated at each iteration, yielding a variable selection path. The optimal thresholding parameter can be chosen by cross-validation. ThrEEBoost was evaluated via simulations to assess the effects of different threshold values on prediction error, sensitivity, specificity, and the number of iterations to identify minimum prediction error under sparse and non-sparse true models with correlated, continuous outcomes. ThrEEBoost requires fewer iterations to locate the coefficients yielding the minimum error. When the true model is sparse, it achieves similar prediction error to an existing non-thresholded technique. When the true model is non-sparse, ThrEEBoost achieves lower prediction error. The technique is illustrated by applying it to the problem of identifying predictors of weight change in a longitudinal nutrition study. An R package implementing the technique, threeboost, is available.

email: *brow3774@umn.edu*

8l. USE OF FUNCTIONAL LINEAR MODELS TO DETECT ASSOCIATIONS BETWEEN CHARACTERISTICS OF WALKING AND HEALTH RELATED OUTCOMES USING ACCELEROMETRY DATA

William F. Fadel*, Indiana University School of Public Health, Indianapolis

Jaroslaw Harezlak, Indiana University School of Public Health, Indianapolis

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

Nancy W. Glynn, University of Pittsburgh

Various methods exist to measure physical activity. Subjective methods, such as diaries and surveys are relatively inexpensive ways of measuring one's physical activity; however, they are riddled with measurement error and bias due to self-report. Wearable accelerometers offer a non-invasive and objective measure of subjects' physical activity and are now widely used in observational studies. Accelerometers record high frequency data and produce an unlabeled time series at the sub-second level. An important activity to identify from the data collected is walking, since it is often the only form of exercise for certain populations. Currently, most methods use an activity summary which ignores nuances of walking data. We propose methodology to model specific health related outcomes (scalar response variable) with a functional linear model utilizing spectra obtained from the local fast Fourier transform (FFT) of walking as a predictor. Walking spectra are transformed from the frequency domain to the order domain so the spectra across subjects are aligned at their average cadence and subsequent harmonics. Utilizing prior knowledge of the mechanics of walking, we incorporate this as additional information of the structure of our transformed walking spectra. Methods are applied to the in-the-lab data obtained from the Developmental Cohort Study (DECOS).

email: *wffadel@iupui.edu*

8m. VARIABLE SELECTION AND COVARIANCE ESTIMATION FOR HIGH DIMENSIONAL DATA

Runmin Shi*, University of Florida

For multivariate regression of big data problems, both variable selection and covariance matrix estimation are usually very hard to handle. In the joint work with my PhD supervisor, Prof. Faming Liang, we developed a method to solve these two problems simultaneously. Like the idea of Gibbs Sampler, we use a consistent estimation, provided the estimated covariance matrix is temporarily fixed, to do the variable selection. After that, we use a consistent estimation of the covariance matrix, based on the updated mean, to update the estimated covariance matrix. Repeat the whole process till some criterions are satisfied. This method has been proved to be asymptotically correct and it can efficiently provide good results in some applications.

email: *shirunmin@foxmail.com*

8n. SCALABLE BAYESIAN VARIABLE SELECTION USING NONLOCAL PRIOR DENSITIES IN ULTRAHIGH-DIMENSIONAL SETTINGS

Minsuk Shin*, Texas A&M University

Anirban Bhattacharya, Texas A&M University

Valen E. Johnson, Texas A&M University

Bayesian model selection procedures based on nonlocal alternative prior densities are extended to ultrahigh dimensional settings and compared to other variable selection procedures using precision-recall curves. Variable selection procedures included in these comparisons include methods based on g-priors, reciprocal lasso, adaptive lasso, scad, and minimax concave penalty criteria. The use of precision-recall curves eliminates the sensitivity of our conclusions to the choice of tuning parameters. We find that Bayesian selection procedures based on nonlocal priors are competitive to all other procedures in a range of simulation scenarios, and we subsequently explain this favorable performance through a theoretical examination of their consistency properties. When certain regularity conditions apply, we demonstrate that the nonlocal procedures are consistent for linear models even when the number of covariates p increases sub-exponentially with the sample size n. Methods based on Zellner's g-prior are also found to be competitive with penalized likelihood methods in identifying the true model, but the posterior distribution on the model space induced by this method is much more dispersed than the posterior distribution induced on the model space by the nonlocal prior methods.

email: *minsuk000@gmail.com*

8o. ON GAUSSIAN COMPARISON INEQUALITY AND ITS APPLICATION TO SPECTRAL ANALYSIS OF LARGE RANDOM MATRICES

Sheng Xu*, Johns Hopkins University

Wenxin Zhou, University of Melbourne

Fang Han, Johns Hopkins University

Very recently, Chernozhukov, Chetverikov, and Kato developed a new Gaussian comparison inequality for approximating the suprema of empirical processes. This paper exploits this technique to devise sharp inference on spectra of large random matrices. In particular, we show how two long-standing problems in random matrix theory can be solved: (1) simple bootstrap inference on sample eigenvalues when true eigenvalues are tied; (2) conducting two-sample Roy's covariance test in high dimensions. A generalized epsilon net argu-

ment regarding the matrix rescaled spectral norm and several new empirical process bounds are developed to establish the asymptotic results, and might be of independent interest.

email: *shxu@jhu.edu*

8p. A UNIFIED THEORY OF CONFIDENCE REGIONS AND TESTING FOR HIGH DIMENSIONAL ESTIMATING EQUATIONS

Matey Neykov*, Princeton University

Yang Ning, Princeton University

Jun S. Liu, Harvard University

Han Liu, Princeton University

We propose a new inferential framework of constructing confidence regions and testing hypotheses for statistical models specified by a system of high dimensional estimating equations. The key ingredient of this framework is an influence function constructed by projecting the fit- ted estimating equations to a sparse direction obtained by solving a large-scale linear program. The main feature of our framework which makes it different from the existing ones is that the specification of the loglikelihood and other types of loss functions is not needed. Our main theoretical contribution is to establish a unified Z-estimation theory of confidence regions for high dimensional problems. In particular, we derive uniformly valid confidence regions for low dimensional parameters of interest. We further apply our general framework to a number of examples including noisy compressed sensing, undirected graphical models, discriminant analysis and vector autoregression models. We provide thorough numerical simulations to back up the developed theoretical results.

email: *mneykov@princeton.edu*

8q. ON THE ESTIMATION OF POPULATION EIGENVALUES AND THE ASYMPTOTIC PROPERTIES OF PCA IN HIGH-DIMENSIONAL DATA

Rounak Dey*, University of Michigan

Seunggeun Lee, University of Michigan

With the development of high-throughput biomedical technologies, principal component analysis (PCA) in high-dimensional regime is of great interest. Existing methods for the estimation of population eigenvalues, eigenvectors, and PC scores are based on a spiked eigenvalue model in which population eigenvalues are one except for a few large eigenvalues. In real data, this assumption may not be satisfied due to the presence of local correlation among features. We propose a novel method to consistently estimate population eigenvalues without the spiked eigenvalue assumption. Our method combines two existing algorithms, one for estimating the large eigenvalues, the other for estimating the distribution of the remaining eigenvalues. Based on the consistent estimator of population eigenvalues, we construct estimators of the angle between sample and population eigenvectors, correlation coefficients between sample and population PC scores, and shrinkage factors of the predicted PC scores. We also provide theoretical justification of the proposed methods using random matrix theory. Extensive simulation studies and real data examples from genetics show the superior performance of our method.

email: *deyrnk@umich.edu*

9. POSTERS: BAYESIAN METHODS AND COMPUTATIONAL ALGORITHMS

9a. INVITED POSTER: BAYESIAN PREDICTIVE MODELING FOR PERSONALIZED TREATMENT SELECTION

Junsheng Ma, 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

Efforts to personalize medicine in oncology have been limited by reductive characterizations of the intrinsically complex underlying biological phenomena. Future advances in personalized medicine will rely on molecular signatures that derive from synthesis of multifarious interdependent molecular quantities requiring robust quantitative methods. However, highly-parameterized statistical models when applied in these settings often require a prohibitively large database and are sensitive to proper characterizations of the treatment-by-covariate interactions, which in practice are difficult to specify and may be limited by generalized linear models. In this paper, we present a Bayesian predictive framework that enables the integration of a high-dimensional set of genomic features with clinical responses and treatment histories of historical patients, providing a probabilistic basis for using the clinical and molecular information to personalize therapy for future patients. Our work represents one of the first attempts to define personalized treatment assignment rules based on large-scale genomic data. We use actual gene expression data acquired from the Cancer Genome Atlas in the settings of leukemia and glioma to explore the statistical properties of our proposed Bayesian approach for personalizing treatment selection. The method is shown to yield considerable improvements in predictive accuracy when compared to penalized regression approaches.

email: *bphobbs@mdanderson.org*

9b. LOGISTIC 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 logistic regression model logit(Pr(Y = 1|X)) = Xβ from large studies, and we would like to utilize this summary information for improving inference in an expanded model of interest, logit(Pr(Y = 1|X, B)) = (X, B)γ, in a new dataset of moderate size. By using logistic regression approximation proposed by Monahan and Stefanski, 1992, we formulate the problem into an inferential framework where the historical information is translated in terms of a set of non-linear constraints on the parameter space. We propose several frequentist and Bayes solutions. For Bayes solutions, these non-linear constraints are treated as informative priors for β and weakly informative Cauchy priors for γ (Gelman et al, 2008). 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 situation. Our simulation results comparing these solutions indicate that historical information on model logit(Pr(Y = 1|X)) = Xβ can boost the efficiency of estimation and enhance prediction ability in the model of interest logit(Pr(Y = 1|X, B)) = (X, B)γ.

email: *chengwt@umich.edu*

9c. A LOW INFORMATION PRIOR SPECIFICATION FOR A DIRICHLET PROCESS MIXTURE OF GAUSSIAN DISTRIBUTIONS

Michael Martens*, Medical College of Wisconsin

Purushottam Laud, Medical College of Wisconsin

A Dirichlet process mixture of Gaussian distributions can model the distributions of scalar and vector valued data extremely closely, given a well-chosen prior. Prior specification is challenging, however; a prior that performs well with one data set will give poor results for others. With each new data set, the user is burdened with the task of finding a prior that provides reasonable modeling. We introduce a robust Dirichlet process mixture of Gaussian distributions that can accurately model a wide variety of distributions. The model requires the user only to provide prior guesses of the median and 95thpercentile of the sampled distribution. Using these, we transform the data to a standardized scale. Then, a Dirichlet process mixture with a low information prior on this scale is applied to this transformed data. Lastly, the fitted model on the transformed data is converted to a model on the original data by back-transformation. Using several simulated datasets, we show that our method provides accurate modeling of a diverse collection of distributions that includes skewed, multimodal, and highly dispersed members.

email: *mmartens@mcw.edu*

9d. SOME EXAMPLES OF BAYESIAN NETWORK META-ANALYSIS OF LONGITUDINAL DATA

Jonathon J. Vallejo*, Baylor University

Network meta-analysis of longitudinal data allows one to account for trends over time in making comparisons among treatments, potentially removing bias and allowing one to compare treatments at times which were not in the designs of some of the studies. In 2009, Jones et al. reviewed the literature for methods of meta-analyzing longitudinal data. Their conclusion was that practitioners were undecided on how to appropriately meta-analyze longitudinal studies, and subsequently Jones et. al proposed a few methods for doing so. Since then, various new methods have been proposed for the meta-analysis of longitudinal data, though few of these explicitly mention longitudinal data. In addition to the research which explicitly references longitudinal data, there also exist developments in the areas of meta-analysis for repeated measures, model-based meta-analysis, and multivariate meta-analysis. Thus, though models have been created and tested in each of these areas, there has been no exploration of the performance of these models against each other in various longitudinal settings. Furthermore, due to this lack of exploration and a cohesive framework, practitioners may still feel unclear in how to proceed with a network meta-analysis of longitudinal data. We seek to address these issues in this presentation.

email: *jonathon_vallejo@baylor.edu*

9e. FREQUENTIST AND BAYESIAN APPROACHES TO THE EVALUATION OF BINARY CLASSIFIERS

Fridtjof Thomas*, University of Tennessee Health Science Center

Binary classifiers are used to classify patients, cells, or any other elements into two groups based on a classification rule. In medical testing and prediction, these classification rules are often based on predictive probabilities for having a certain condition obtained by logistic regression with risk factors and other covariates as independent variables, but other approaches like classification trees or neural networks exist as well. Competing models are frequently compared based on the resulting receiver operating curve (ROC) characteristics and the associated area under the ROC curve (AUC). Many statistical programs readily print confidence intervals for the AUC, but AUC itself does not evaluate a single binary classifier but instead summarizes the performance of all possible binary classifiers that might be derived from a given model where each binary classifier results from a specific threshold for classification with associated sensitivity (true positive rate) and specificity (true negative rate). We contrast uni- and multivariate frequentist confidence intervals and Bayesian posterior density intervals for a number of associated quantities of interest such as: the threshold

to achieve a pre-specified false positive rate; the sensitivity and specificity associated with a given threshold; and the prevalence dependent positive and negative predictive values.

email: *fthomas4@uthsc.edu*

9f. SPATIO-TEMPORAL BAYESIAN QUANTILE REGRESSION FOR ANALYZING WEATHER DATA OF US

Priyam Das*, North Carolina State University

Subhashis Ghoshal, North Carolina State University

We consider a Bayesian method for simultaneous quantile regression on a real variable. By monotone transformation, we can make both the response variable and the predictor variables in the unit interval. A representation of quantile function is given by a convex combination of two monotone increasing functions not depending on the prediction variables. In a Bayesian approach, a prior is put on quantile functions by putting prior distribution those monotone functions independently. The monotonicity constraint on the curves are obtained through a spline basis expansion with coefficients increasing and lying in the unit interval. We put a Dirichlet prior distribution on the spacings of the coefficient vector. A finite random series based on splines obeys the shape restrictions. Taking the tensor product of B-spline basis function to account for the space variability of the response variable, we propose a novel method for spatiotemporal quantile regression for two or more dimensional space. We apply our method to analyze the weather data of USA.

email: *pdas@ncsu.edu*

9g. SIMULATION-BASED ESTIMATION OF MEAN AND STANDARD DEVIATION FOR META-ANALYSIS USING APPROXIMATE BAYESIAN COMPUTATION (ABC) COUPLED WITH MODEL AVERAGING METHOD

Deukwoo Kwon, University of Miami

Isildinha M. Reis*, University of Miami

When conducting a meta-analysis of a continuous outcome, estimated means and standard deviations from the selected studies are required. If these quantities are not directly reported in the publications, they must be estimated from other reported summary statistics, such as the median, the minimum, the maximum, and quartiles. In our previous publication, we proposed a simulation-based estimation approach using the Approximate Bayesian Computation (ABC) technique for estimating mean and standard deviation based on various sets of summary statistics found in published studies. Our approach outperformed the other available methods when data are generated from skewed or heavy-tailed distributions. The previous ABC with single distribution selection was based on assumed single parametric distribution of original data. We need to choose one distribution from several candidate distributions, using posterior model probability. We found that the selection of an underlying distribution via posterior model probability was sensitive to the prior distribution for parameters. In this study, we exploit the use of ABC with model averaging methodology to estimate mean and standard deviation. We show that ABC coupled with model averaging of several candidate distributions performs better than the ABC with distribution selection in terms of the average relative errors (AREs).

email: *ireis@miami.edu*

9h. A LOW INFORMATION PRIOR FOR DIRICHLET PROCESS MIXTURE OF WEIBULL DISTRIBUTIONS

Yushu Shi*, Medical College of Wisconsin

Purushottam Laud, Medical College of Wisconsin

In 2008, Kottas proposed a Dirichlet process mixture (DPM) of Weibull distributions as a model for survival data that performs well given appropriate priors. However, the choice of priors for DPM models that is suitable for a wide variety of data situations is somewhat elusive. With some minor modification to Kottas' model, we develop a simple scheme which only requires the user to specify a prior guess at a high percentile of the population's distribution. This value is used to transform the data to a standard scale on which a low information prior is constructed. After drawing samples from the posterior with the scaled data, we rescale the inference back to the original units. The low information prior is selected to provide a wide variety of Weibull components for the DPM in order to generate flexible distributions for the data on the standard scale. We also extend Kottas' model include interval censored data, and to the case with covariates. With simulated scenarios we demonstrate that the method gives satisfying answers under challenging situations with different types of censoring and data generating distributions.

email: *yushushi@mcw.edu*

9i. MCMC METHODS FOR BAYESIAN MODEL SELECTION FOR LOG-BINOMIAL REGRESSION

Wei Zhou*, University of Cincinnati

Siva Sivaganesan, University of Cincinnati

In epidemiological and clinical studies, relative risk (RR) is often the preferred measure of exposure effect. Log-binomial regression can be used to model RR, as its coefficients naturally offer RR. However, the constrained parameter space of the log-binomial model often causes convergence failures for standard algorithms to locate the maximum likelihood estimate. Frequentist methods had been proposed to resolve this issue but they are often not reliable. Furthermore, there is little literature discussion about the variable selection for log-binomial regression, due to its intractable constrained parameter space. Bayesian approach can deal more easily with such space and is intuitive to implement, therefore is a

viable alternative for log-binomial modeling. Our research performs Bayesian variable selection for the log-binomial model through the Bayes factor, a central tool in the Bayesian hypothesis testing theory. Markov Chain Monte Carlo (MCMC) methods are used for Bayes factor estimation. We survey and evaluate the following MCMC methods: harmonic mean method, importance sampling, reciprocal importance sampling, Carlin and Chib's method and bridge sampling. The performances of these methods are compared on simulated data as well as real data. Slice sampler is used for posterior sampling.

email: *zhouwei0824@gmail.com*

9j. A BAYESIAN HIERARCHICAL SUMMARY RECEIVER OPERATING CHARACTERISTIC MODEL FOR NETWORK META-ANALYSIS OF DIAGNOSTIC TESTS

Qinshu Lian*, University of Minnesota

Haitao Chu, University of Minnesota

In studies evaluating the accuracy of diagnostic tests, three designs are commonly used: (1) the crossover design; (2) the randomized design; and (3) the non-comparative design. Existing methods on meta-analysis of diagnostic tests mainly considered the simple cases when the reference test in all or none of the studies can be considered as a gold standard test, and when all studies use either a randomized or non-comparative design. Yet the proliferation of diagnostic instruments and diversity of study designs being used have boosted the demand to develop more general methods to combine studies with or without a gold standard test using different designs. In this paper, we extend the Bayesian hierarchical summary receiver operating characteristic model to network meta-analysis of diagnostic tests to simultaneously compare multiple tests under a missing data framework. It accounts for the potential correlations between multiple tests within a study and the heterogeneity across studies. In addition, it allows different studies to perform different subsets of diagnostic tests and provides flexibility on the choice of summary statistics. Our model is evaluated through simulations and illustrated using real data from deep vein thrombosis tests.

email: *lianx025@umn.edu*

9k. PRIOR ELICITATION VIA A RORSCHACH-STYLE GRAPHICAL PROCEDURE

Christopher Casement*, Baylor University

David Kahle, Baylor University

Prior specification is fundamental to the Bayesian paradigm. Informative priors allow analysts to incorporate expert opinion directly into the modeling process. However, using such priors can strongly influence an analysis, so using informative priors demands a principled approach to prior specification. Prior elicitation allows the quantification of an expert's belief into a probability distribution. When eliciting an informative prior, standard methods involve facilitators asking experts to quantify their beliefs in the form of multiple distribution summaries, such as means, modes, and specific percentiles, with the statistician then converting these back into the standard parameters of a given family. While software exists that assists experts in the process, eliciting a prior distribution that accurately reflects expert opinion is still a challenging process. In this work we propose an interactive, visualization-based tool that enables prior specification without the expert needing to explicitly quantify her beliefs; rather, she passes through a series of Rorschach-style tests where she selects the dataset that she believes to be most likely, and the algorithm does the rest. To illustrate the method's ability to accurately quantify expert opinion, we consider the prior elicitation of a population proportion. The process is implemented in a user-friendly Shiny application.

email: *chris_casement@baylor.edu*

9l. BREGMAN DIVERGENCE TO GENERALIZE BAYESIAN INFLUENCE MEASURES FOR DATA ANALYSIS

Matthew M. Weber*, Florida State University

Debajyoti Sinha, Florida State University

Dipak K. Dey, University of Connecticut

This paper introduces and demonstrates the use of Bregman divergence measures for generalizing and extending existing popular Bayesian influence diagnostics. We derive useful properties of these Bregman divergence based cross-validated measures of influential observations. We show that these cross-validated Bregman divergence based influence measures can be computed via Monte Carlo Markov Chain samples from a single posterior based on full data. We illustrate how our measures of influence of observations have more useful practical roles for data analysis than popular Bayesian residual analysis tools using a meta-analysis of clinical trials under generalized linear models.

email: *mweber@stat.fsu.edu*

9m. A BAYESIAN SCREENING APPROACH FOR HEPATOCELLULAR CARCINOMA USING TWO LONGITUDINAL BIOMARKERS

Nabihah Tayob*, University of Texas MD Anderson Cancer Center

Francesco Stingo, 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

Advanced hepatocellular carcinoma (HCC) has limited treatment options and poor survival. Early detection of HCC is critical to improve the prognosis of these patients. Current guidelines for high-risk patients include six-month ultrasound screenings but these are not sensitive for early HCC. Alpha-fetoprotein (AFP) is a widely used diagnostic biomarker but has shown limited use in HCC screening with a fixed threshold. Approaches that incorporate longitudinal AFP have shown potentially increased detection of HCC but AFP is not elevated in all HCC cases so we incorporate a second HCC biomarker, des-gamma-carboxy prothrombin (DCP). The data from the Hepatitis C Antiviral Long-term Treatment against Cirrhosis (HALT-C) Trial is a valuable source of data to study biomarker screening. We assume the trajectories of AFP and DCP follow a joint hierarchical mixture model with random change points. Markov chain Monte Carlo methods are used to calculate posterior distributions used in risk calculations among future patients. The posterior risk of HCC, given longitudinal values of AFP and DCP, is used to determine whether a patient has a positive screen. The screening algorithm was compared to alternatives in the HALT-C Trial (using cross-validation) and in simulations studies under a variety of possible scenarios.

email: *ntayob@gmail.com*

9n. BAYESIAN SINGLE INDEX MODELS

Kumaresh Dhara*, Florida State University

Single index model is a popular tool in a wide variety of applications in biomedical research. From a frequentist viewpoint, there is a substantial amount of literature devoted to proposing strategies for estimating and inferring parameters of the model, with corresponding theoretical justification regarding consistency and asymptotic efficiency. On the other hand, there are currently only a handful articles on Bayesian single index models. In a Bayesian setup, we need a suitable prior process for the unknown nonparametric univariate function and another prior distribution on the sphere for the unknown coefficients. While using these types of priors, the posterior distribution of the coefficients does not have a closed form expression, and existing tools typically results in slow mixing of the Markov chain. In this article, we propose an efficient variant of the Metropolis Hastings algorithm to sample from the full conditional distribution using an Orsntein Uhlenbeck (OH) process for the nonparametric effect and a carefully chosen proposal density based on model alignment. The use of OU process allows efficient evaluation of the likelihood and results in fast convergence of the Markov chain.

email: *k.dhara@stat.fsu.edu*

10. POSTERS: SEMI- AND NON-PARAMETRIC METHODS

10a. ON THE ASYMPTOTIC DISTRIBUTION OF THE WILCOXON SIGNED RANK TEST STATISTIC

Xueyi Chen*, University of Kansas Medical Center

Francisco J. Diaz, University of Kansas Medical Center

The Wilcoxon signed rank test statistic (T^+) is widely employed in one sample nonparametric tests regarding the sample median, assuming only that the underlying distribution is symmetric. It uses the sign of the differences D_i between observations and the median, under the null hypothesis, and the magnitude of these observations. When the sample size is large, the distribution of this statistic can be approximated by a normal distribution. To date authors such as Gibbons et. al. [1] have outlined the calculation of the mean and variance of the Wilcoxon signed rank test statistic, however detailed justification of independence between the sign indicator Z_i and the rank of D_i, $r(D_i)$ has typically not been provided. Authors have also assumed that $T^+=\Sigma\,\Sigma\,1_{\leq i\leq j\leq N}\,T_{ij}$ in the computation of the variance of T^+ without rigorous mathematical proof, where T_{ij} is an indicator of the sign of $D_i + D_j$ and N is the sample size . In this work, we fill in the details for the justification of the asymptotic distribution of Wilcoxon signed rank test statistic, presenting two proofs: (1) The proof of independence between Z_i and $r(D_i)$ with the assumption that the distribution is symmetric; and (2) the proof that in general: $T^+=\Sigma\,\Sigma\,1_{\leq i\leq j\leq N}\,T_{ij}$ [1] Gibbons, J. D. and Chakraborti, S., Nonparametric Statistical Inference 5th ed., Taylor & Francis Group, New York, 2011.

email: *xchen@kumc.edu*

10b. SEMIPARAMETRIC SURVIVAL MODEL WITH TIME-DEPENDENT CURE PROCESS

Sophie Yu-Pu Chen*, University of Michigan

Alexander Tsodikov, University of Michigan

Cure models refer to survival models incorporating a cure fraction. As medical treatments progress, cure models has been applied to time-to-event data for diseases where a proportion of patients are at no risk of the disease following treatment. In most of the current work on cure models it is assumed that the cure status is determined, if unknown, at the beginning of the follow up (t=0). In practice though, patients often receive treatments during the follow up. In this case it is natural to expect the chance of cure to change over time in response to treatment. To account for this situation, we propose a joint dynamic model for the cure process and a terminal event. Two separate baseline hazards are estimated nonparametri-

cally in the model to allow different time scales for the cure and time to failure processes. An EM algorithm is developed to estimate the two infinite dimensional parameters. Covariates are modeled parametrically and estimated using the profile likelihood. Large-sample properties are obtained. Simulation studies are presented to illustrate the finite-sample properties. The proposed model is applied to the prostate cancer data from the SEER program.

email: *yupuchen@umich.edu*

10c. SEMIPARAMETRIC MODELS OF BIVARIATE TIMES TO EVENT DATA WITH A SEMICOMPETING RISK

Ran Liao*, Indiana University, Bloomington

Sujuan Gao, Indiana University, Indianapolis

Survival analysis of time to events data often encounters the situations of correlated multiple events including the same type of event observed from siblings or multiple events experienced by the same individual. In addition, survival analysis in biomedical research can be further complicated by semi-competing risk when individuals at risk of a particular disease die from other causes. In this poster, we propose a frailty model based approach for bivariate survival outcomes with a semi-competing risk. Two estimation approaches are proposed and compared. The first is a two-stage semiparametric approach where the cumulative baseline hazard was estimated by a nonparametric method first and plugged in the likelihood function. Parameter estimation was then achieved by maximizing the pseudo-likelihood functions. In the second approach, we propose to use a pseudo partial likelihood approach for parameter estimation and inference similar to the concept in the Cox's partial likelihood. Simulation studies are conducted to compare the performances of these two approaches. The proposed model is applied to data from a longitudinal study of an elderly population.

email: *ranliao@iu.edu*

10d. CHANGE-PLANE ANALYSIS FOR SUBGROUP DETECTION AND SAMPLE SIZE CALCULATION

Ailin Fan*, North Carolina State University

Rui Song, North Carolina State University

Wenbin Lu, North Carolina State University

We propose a systematic method for testing and identifying a subgroup with an enhanced treatment effect. We adopts a change-plane technique to first test the existence of a subgroup, and then identify the subgroup if the null hypothesis on non-existence of such a subgroup is rejected. A semiparametric model is considered for the response with an unspecified baseline function and an interaction between a subgroup indicator and treatment. A doubly-robust test statistic is constructed based on this model, and asymptotic distributions of the test statistic under both null and local alternative hypotheses are derived. Moreover, a sample size calculation method for subgroup detection is developed based on the proposed statistic. The finite sample performance of the proposed test is evaluated via simulations. Finally, the proposed methods for subgroup identification and sample size calculation are applied to a data from an AIDS study.

email: *afan@ncsu.edu*

10e. APPROXIMATING SMALL P-VALUES IN PERMUTATION TESTS: USING THE STRUCTURE OF THE PERMUTATION SPACE TO SPEED UP COMPUTATION

Brian D. Segal*, University of Michigan

Hui Jiang, University of Michigan

Thomas Braun, University of Michigan

Researchers in genetics and other life sciences commonly use permutation tests to evaluate differences between groups. Permutation tests have desirable properties, including exactness, and are applicable even when the distribution of the test statistic is analytically intractable. However, permutation tests can be computationally intensive. We propose an algorithm for quickly approximating small permutation p-values in two-sample tests. Our approach is based on a stochastic ordering of test statistics across partitions of the permutation space, which allows us to calculate p-values in partitions that require less computation and then predict p-values in partitions that would require more computation. In this article, we present our method and demonstrate its use through simulations and an application to cancer genomic data. We find that our method is faster than a current leading method, and can successfully identify up- and down-regulated genes.

email: *bdsegal@umich.edu*

10f. WEIGHTED SEMI-PARAMETRIC REGRESSION MODELS FOR DOUBLY TRUNCATED SURVIVAL DATA

Lior Rennert*, University of Pennsylvania

Sharon X. Xie, University of Pennsylvania

Double truncation often arises in survival data, when the data is only observed if it falls within a particular time interval. If we do not account for double truncation, estimators of regression coefficients in the standard Cox regression model will be biased. In this paper, we propose a weighted semi-parametric regression model for doubly truncated data. Here the weights correspond to the probability of a particular subject being observed, and are estimated non-parametrically. Through extensive simulations, we show that the weighted regression coefficient estimator is unbiased. Furthermore, in many situations the weighted regression coefficient estimator has a smaller mean square error than the estimator from the stan-

dard Cox model. We show that our proposed estimator is consistent and asymptotically normal. We illustrate the proposed method using a mental health data set.

email: *lior.rennert@gmail.com*

10g. NON-PARAMERIC SHRINKAGE MEDIAN ESTIMATION

Beidi Qiang*, University of South Carolina

Edsel Pena, University of South Carolina

The problem of finding shrinkage estimators which dominates the usual sample mean estimator has been well studied in literature. Most existing methods are derived based on normal distribution. Some nonparametric shrinkage methods are studied, but those are still under the assumption of finite mean and variance. In this study, a shrinkage estimator for the population median is proposed. The new estimator does not assume a specific parametric distribution and it does not require the existence of finite moments. The practical improvement of the proposed estimator is demonstrated through simulation studies and data analysis. The proposed method performs better than the usual estimates in heavy tailed distributions especially when sample size is small.

email: *Qiangb@email.sc.edu*

10h. ROBUST NONPARAMETRIC KERNEL REGRESSION ESTIMATOR

Ge Zhao*, University of South Carolina

Yanyuan Ma, University of South Carolina

We develop a robust nonparametric kernel regression estimator. The estimator controls the effect from outlying observations through a combination of weighting and trimming. The estimator is obtained through solving estimating equations, where a trimming parameter and a bandwidth are needed as tuning parameters. We prescribe a data driven method to select these tuning parameters in the spirit of cross-validation. We also show asymptotic consistency, establish the estimation bias, variance properties and derive the asymptotic distribution of the resulting estimator. The finite sample performance of the estimator is illustrated through both simulation studies and analysis on a problem related to wind power generation, which motivated this study at the first place.

email: *zichuan1028@gmail.com*

10i. A RANDOM FOREST OF MODIFIED INTERACTION TREES FOR TREATMENT DECISION RULES

Zhen Zeng*, Merck

Wei Zheng, Sanofi

Yuefeng Lu, Sanofi

Personalized medicine, or precision medicine, is a medical model that uses disease subtypes, genetic makers, and other patient-level factors to develop customized treatment with desirable benefit/risk profiles for a given patient. In recent years, various statistical methodologies have been developed in this domain but there are still many open questions. We propose a novel tree-based ensemble method, a random forest of modified interaction trees (RFMIT), to generate predictive importance scores for covariates, and directly predict treatment effects for each individual patient with confidence intervals. This method can be used to select predictive biomarkers, visualize treatment effects, generate predictive models, and easily incorporate a clinically meaningful difference for treatment decision or future enrichment design.

email: *zhenhouse@msn.com*

10j. A PROFILE MAXIMUM PSEUDOLIKELIHOOD ESTIMATOR FOR THE PROPORTIONAL CAUSE-SPECIFIC HAZARDS MODEL UNDER OUTCOME MISCLASSIFICATION

Giorgos Bakoyannis*, Indiana University School of Medicine and Richard M. Fairbanks School of Public Health

Ying Zhang, Indiana University School of Medicine and Richard M. Fairbanks School of Public Health

Constantin T. Yiannoutsos, Indiana University School of Medicine and Richard M. Fairbanks School of Public Health

Competing risks data arise naturally in observational cohort studies and clinical trials. Frequently, outcome determination is based on imperfect, and usually less expensive, diagnostic procedures leading to errors in outcome classification. This type of misclassification can lead to seriously biased estimates and thus invalid conclusions. In such cases, we propose a double sampling design to retrieve the true outcome for a small sample of non-censored cases, by using a gold standard and possibly more expensive diagnostic procedure. Based on this information we estimate the probabilities of the true outcomes conditional on the imperfect diagnosis and other covariates, and plug these probabilities in the profile likelihood function to obtain a profile pseudolikelihood. We show that the corresponding profile maximum pseudolikelihood estimator (PMSLE) is consistent and asymptotically normal. Simulation studies show that the naive estimator is highly biased under outcome misclassification, whereas the PMSLE works well with small sample sizes and is fairly robust against misspecification of the model for the true outcome probabilities. The method is illustrated using data from HIV-1 seropositive individuals in sub-Saharan Africa, where serious death under-reporting results in classifying many deceased patients as being disengagers from HIV care.

email: *gbakogia@iu.edu*

10k. LIKELIHOOD RATIO TESTING IN FUNCTIONAL ADDITIVE MODELS

Merve Yasemin Tekbudak*, North Carolina State University
Marcela Alfaro-Cordoba, North Carolina State University
Ana-Maria Staicu, North Carolina State University
Arnab Maity, North Carolina State University

A functional additive model is a regression model of a scalar or functional response on a finite number of functional principal component scores of the functional covariate. Using a mixed model representation, we consider the exact likelihood and restricted likelihood ratio tests for testing the nullity and linearity of the effect of the functional covariate in the context of scalar-on-function additive models. We assume that the functional covariate is observed on a dense or sparse grid and without measurement error. An extensive simulation study is performed to investigate Type I error rate and power, of both the likelihood ratio test (LRT) and restricted likelihood ratio test (RLRT). Their performances are compared with all other alternative approaches available in the literature.

email: *mytekbud@ncsu.edu*

10l. NONPARAMETRIC CHANGE POINT DETECTION METHODS FOR PROFILE VARIABILITY

Vladimir J. Geneus*, Florida State University

A wavelet-based change point method is proposed that determines when the variability of the noise in a sequence of functional profiles (i.e. the precision profile of medical devices) goes out of control from a known, fixed value or an estimated, in-control value. The functional portion of the profiles is allowed to come from a large class of functions and may vary from profile to profile. Our proposed method makes use of the orthogonal properties of wavelet projections to accurately and efficiently monitor the level of noise from one profile to the next. The estimator is evaluated on a variety of conditions, including allowing the wavelet noise subspace to be substantially contaminated by the profile's functional structure, and is compared to two competing noise monitoring methods. The proposed method is shown to be very efficient at detecting when the variability has changed through an extensive simulation study. Extensions are proposed which will explore the non-Gaussian assumption throughout our applications, the usage of windowing and non in-control values for the MAD method, and the effect of the exact distribution under normality rather than the asymptotic distribution. The proposed methodology is tested through simulation and applicable to various biometric and health related topics.

email: *vgeneus@stat.fsu.edu*

10m. COVARIATE ADJUSTED SPEARMAN'S RANK CORRELATION WITH PROBABILITY-SCALE RESIDUALS

Qi Liu*, Vanderbilt University
Bryan Shepherd, Vanderbilt University
Valentine Wanga, Institute for Health Metrics and Evaluation
Chun Li, Case Western Reserve University

It is desirable to adjust Spearman's rank correlation for covariates, yet existing proposals have limitations. We propose two estimators for covariate-adjusted Spearman's rank correlations, partial and conditional, using probability-scale residuals (PSRs). Our partial estimator is the correlation of PSRs from models of X on Z and of Y on Z, which is elegantly analogous to the partial Pearson's correlation derived as the correlation of observed-minus-expected residuals. Our conditional estimator is the conditional correlation of PSRs, which can be used to capture changes in Spearman's rank correlations between different values of covariates. Our estimators are very general, applicable to any orderable variable. With PSRs obtained from semiparametric transformation models, our estimators preserve the rank-based nature of Spearman's rank correlation. We conduct simulations to evaluate the performance of our estimators and compare them with other popular measures of association, demonstrating their robustness and efficiency. Our method is illustrated in two application examples: one looking at the association between workers' educational attainments and their wages in the United States after controlling for other potentially confounding variables, and a second application estimating pairwise correlations between responses to all questions in a large survey performed in Mozambique after adjusting for relevant demographic and community-level factors.

email: *qi.liu.1@vanderbilt.edu*

11. POSTERS: CENSORING, TRUNCATION, AND MISSINGNESS

11a. TRUNCATION-BASED NEAREST NEIGHBORS IMPUTATION FOR HIGH DIMENSIONAL DATA WITH DETECTION LIMIT THRESHOLDS

Jasmit S. Shah*, University of Louisville
Guy N. Brock, University of Louisville
Shesh N. Rai, University of Louisville
Aruni Bhatnagar, University of Louisville

High throughput technology makes it possible to monitor metabolites on different experiments and has been widely used to detect differences in metabolites in many areas of biomedical research. Mass spectrometry has become one of the main analytical

techniques for profiling a wide array of compounds in the biological samples. Missing values in metabolomics dataset occur widely and can arise from different sources, including both technical and biological reasons. Mostly the missing value is substituted by the minimum value, and this substitute may lead to different results in the downstream analyses. In this study we propose a modified version of the K-nearest neighbor (KNN) approach which accounts for the truncation at the minimum value called KNN truncation (KNN-TN). We compare the imputation results based on KNN-TN with other KNN approaches such as KNN based on correlation (KNN-CR) and KNN based on Euclidean distance (KNN-EU). The proposed approach assumes that the data follows a truncated normal distribution with the truncation point at the detection limit (LOD). The mean and standard deviation of each metabolite are estimated assuming the data arise from a truncated sample, and these estimates are used in the standard KNN-CR algorithm. The results of KNN-TN, KNN-CR and KNN-EU were rigorously tested to estimate the missing values in different types of datasets and their effectiveness was analyzed by the root mean square error (RMSE) measure. The results reported that KNN-TN showed improved performance in imputing the missing values of the different datasets compared to KNN-CR and KNN-EU.

email: *jasmit.shah@louisville.edu*

11b. MULTIPLE IMPUTATION OF MISSING COVARIATES FOR THE COX PROPORTIONAL HAZARDS CURE MODEL

Lauren J. Beesley*, University of Michigan

Jonathan W. Bartlett, London School of Hygiene & Tropical Medicine

Jeremy M. G. Taylor, University of Michigan

We explore several approaches for imputing partially observed covariates when the outcome of interest is a censored event time and when there is an underlying subset of the population that will never experience the event of interest. We call these subjects "cured" and we consider the case where the data is modeled using a Cox Proportional Hazards (CPH) mixture cure model. We study covariate imputation approaches using fully conditional specification (FCS). We derive the exact full conditional distribution and suggest a sampling scheme for imputing partially observed covariates in the CPH cure model setting. We also propose several approximations to the exact distribution that are simpler and more convenient to use for imputation. A simulation study demonstrates that the proposed imputation approaches outperform existing imputation approaches for survival data without a cured fraction in terms of bias in estimating CPH cure model parameters. We apply our multiple imputation techniques to a study of previously untreated patients with head and neck cancer (HNSCC).

email: *lbeesley@umich.edu*

11c. SEQUENTIAL BART FOR IMPUTATION OF MISSING COVARIATES

Dandan Xu*, University of Florida

Michael J. Daniels, University of Texas, Austin

Almut G. Winterstein, University of Florida

To conduct comparative effectiveness research using electronic health records (EHR), many covariates are typically needed to adjust for selection and confounding biases. Unfortunately, it is typical to have missingness in these covariates. Just using cases with complete covariates will result in considerable efficiency losses and likely bias. Here, we consider the covariates missing at random (MAR) with missing data mechanism (MDM) either depending on the response or not. Standard methods for multiple imputation can either fail to capture nonlinear relationships or suffer from the incompatibility and uncongeniality issues. We explore a flexible Bayesian nonparametric approach to impute the missing covariates which involves factoring the joint distribution of the covariates with missingness into a set of sequential conditionals and applying Bayesian additive regression trees (BART) to model each of these univariate conditionals. Using data augmentation, the posterior for each conditional can be sampled simultaneously. We provide details on the computational algorithm and make comparisons to other methods, including parametric sequential imputation and two versions of multiple imputation by chained equations (MICE). We illustrate the proposed approach on EHR data from an affiliated tertiary care institution to examine factors related to hyperglycemia.

email: *dxu@cop.ufl.edu*

11d. SAMPLING METHODS TO IMPROVE THE EFFICIENCY OF TWO-PHASE ESTIMATORS FOR A CONTINUOUS OUTCOME

Paul M. Imbriano*, University of Michigan

Trivellore E. Raghunathan, University of Michigan

A two-phase survey design is typically used when a single outcome of interest Y is expensive to obtain, but a surrogate variable X can be measured cheaper. The first phase takes a random sample from the population and measures X, and the second phase uses a subsample from phase one to measure Y. When Y is continuous, a regression estimator is used to estimate μ_y, the mean of Y. Using simple random sampling for phase two does not reduce the marginal variance of our estimate for μ_y. We propose a new method of selecting phase two samples so that the magnitude of the difference in the mean of X between those included in phase two and those excluded is bounded. By restricting the difference in means, we can achieve a large reduction in the marginal variance of our estimate of μ_y, and the variance and reduction can be easily estimated for any preselected bound. The variance reduction over

simple random sampling depends only on the correlation of X and Y and the bound. When correlation is high we can achieve over 50% reduction in variance. The accuracy of our variance estimate was confirmed through simulation.

email: *pimbri@umich.edu*

11e. BAYESIAN NONPARAMETRIC FEATURE SELECTION OVER LARGE-SCALE GENE NETWORKS WITH MISSING VALUES

Zhuxuan Jin*, Emory University

Zhou Lan, North Carolina State University

Jian Kang, University of Michigan

Tianwei Yu, Emory University

In the analysis of high-throughput data, feature selection over large-scale gene networks has become increasingly important. It can provide useful information to facilitate the development of pathology and biology. Most existing methods focus on selecting important genes/sub-networks without distinguishing specific types of effects. Also, one common strategy for the incomplete data problem is to simply remove the unmeasured nodes from the network structure, which may lose the selection accuracy and introduce bias in estimating the gene effects. To address those limitations, in this work, we propose a Bayesian nonparametric method for gene and sub-network selection. It can identify important genes with two different behaviors: "down-regulated" and "up-regulated" for which a novel prior model is developed for the selection indicator incorporating the network dependence. In posterior computation, we resort to Swendsen-Wang algorithm for efficiently updating selection indicators and develop both fully Bayesian inference algorithm and fast approximate Bayesian inference algorithm. The proposed method can straightforwardly take into account missing data, which improves the selection accuracy and reduces the bias in estimating gene effects. We illustrate our methods on simulation studies and the analysis of the gene expression in primary acute lymphoblastic leukemia (ALL) associated with methotrexate (MTX) treatment.

email: *zjin23@emory.edu*

11f. BINARY EXPOSURE AND LONGITUDINAL COGNITION OUTCOMES IN THE PRESENCE OF NON-INGORABLE DROPOUT AND DEATH

Maria Josefsson*, Umea University

Xavier de Luna, Umea University

Michael J. Daniels, University of Texas, Austin

Lars Nyberg, Umea University

In this study we want to investigate the association between losing a partner and cognitive decline, using data from a large population based study where participants are followed over 15 years. One complication is that longitudinal data of older cohorts are prone to dropout due to illness, e.g. dementia, or death. Hence, the non-response may be directly related to the cognitive outcome of interest in which case it is not ignorable. An additional consequence of the existing dropout, is that, not only the outcome, but information regarding family status may also be missing, and the exposure is not observed. In the presence of such incomplete data scenarios, conventional statistical methods are invalid and may lead to biased estimates. We therefore develop a modeling strategy within a Bayesian framework to deal with these issues, and provide inference on the effect of a (time-varying) binary exposure on a longitudinal outcome in the presence of dropout and death.

email: *maria.josefsson@umu.se*

11g. NONPARAMETRIC IMPUTATION FOR NONIGNORABLE MISSING DATA

Domonique Watson Hodge*, Emory University

Qi Long, Emory University

Missing data is a very common phenomenon in studies across different disciplines. Multiple imputation accounts for the uncertainty about the underlying true values and is a very popular technique due to its ease of use. However, the vast majority of imputation techniques are designed for ignorable missing data since non-ignorability is an assumption more challenging to handle. Under non-ignorable missingness, one assumes the nonresponse mechanism depends on unobserved values, and the outcome model for the variable with missing values and the nonresponse model must be modeled jointly. Consequently, joint modeling can produce results that are sensitive to the misspecification of the two models. We propose a more robust, nonparametric technique to multiply impute missing data in the presence of non-ignorability by allowing users to choose optimal weights so the resulting estimator can rely more heavily on the model that is more likely to be correctly specified. Using the two models, we derive predictive scores to achieve dimension reduction and use the resulting scores coupled with a nearest neighbor hot deck to multiply impute the missing values. The new proposed method is shown to outperform several existing multiple imputation methods for non-ignorable missing data in simulations. In addition, the method is illustrated using a real data example from the Georgia Coverdall Acute Stroke Registry.

email: *domonique.watson@emory.edu*

12. POSTERS: CLASSIFICATION, TESTING, AND NETWORKS

12a. MODELING OVERDISPERSED NUCLEAR BUD COUNT DATA USING THE GENERALIZED MONOTONE INCREMENTAL FORWARD STAGEWISE METHOD

Rebecca Ruffin Lehman*, Virginia Commonwealth University

Colleen Jackson-Cook, Virginia Commonwealth University

Kellie Archer, Virginia Commonwealth University

Nuclear buds (NBuds) are a measure of chromosomal instability indicative of genomic damage. While the process of NBud formation is not fully understood, it is believed that NBuds originate from the elimination of amplified DNA that is still attached to the nucleus by nucleoplasmic material. NBud frequency is determined by counting the number binucleated cells with at least one NBud in approximately 2,000 binucleated cells. To elucidate molecular mechanisms involved in DNA damage, we are interested in identifying genomic features associated with NBuds. When analyzing count data often methods that assume the underlying distribution is Gaussian are inappropriate. Although transformations could be applied it often is of interest to analyze the count data using Poisson or negative binomial regression. Negative binomial regression accommodates rate data (total number of binucleated cells scored by subject) and can handle overdispersion. In high-throughput genomic experiments the number of samples does not exceed the number of explanatory variables thus traditional statistical methods cannot be applied. We present our extension of the Generalized Monotone Incremental Forward Stagewise Method to the negative binomial regression model. Results will be described predicting Nbuds using methylation levels of CpG sites in women with breast cancer.

email: *lehmanrr@vcu.edu*

12b. THE OPTIMAL POINT WHEN INTEREST IS IN ONLY A PORTION OF THE ROC CURVE

Donna K. McClish*, Virginia Commonwealth University

The partial area has become fairly popular as a summary measure of accuracy when interest is not in the entire range of false positive rates (FPRs) or true positive rates (TPRs) along the ROC curve. Suppose we are only interested in the portion of the curve that has FPR no larger than a value FPR0. Basing optimality on the Youden Index (YI), formulae are available to calculate the "global" optimal point c*. If FPR(c*)<=FPR0then our global optimal point provides an FPR that is acceptable and would be usable clinically. Otherwise, if FPR(c*)>FPR0then we could consider, instead, a point that would provide the largest YI over the restricted set of FPRs of interest. Generally, this optimal point will be a value larger than c* (assuming high values are associated with disease). In a similar fashion, researchers/clinicians may want a threshold which would guarantee that TPR(c*)>=TPR0. An optimal point can be found that is restricted to the range of TPRs of interest. We present appropriate values for the restricted optimal point, as well as guidance on how often the global and restricted optimal points may diverge.

email: *mcclish@vcu.edu*

12c. AUTOMATION OF IMMUNO-ONCOLOGY FLOW CYTOMETRY ASSAY USING CASK-CYTO

Shubing Wang*, Merck

Junshui Ma, Merck

David Alexander Merck

George Skibinski, Merck

Jinkai Teo, Merck

Janice Hsueh Ling Oh, Merck

Due to the tremendous success of Keytuda, immunophenotypic analysis of single cells has become strategically important to Immune-Modulatory Receptor (IMR) discovery at Merck. Being able to simultaneously analyze up to multiple intra-cellular and surface markers, including a set of key IMRs, Flow Cytometry (FCM) Immuno-Oncology assays address immunophenotyping with extended and higher-dimensional single-cell analysis. High-dimensional FCM data analysis, though extremely attractive in discovery, imposes scalability and subjectivity issues on standard manual gating practice, therefore lack of analytical power for thorough, identifiable correlation-based mining and statistical inference. We developed a Customized Advanced but Simple Kits for Cytometry (CASK-Cyto) data analysis platform to automatically identify all the cellular subpopulations using a data-driven hierarchical clustering algorithm, and as a proof-of-concept, we applied this method to automatically identify target cell populations and functional cell markers in multiple immuno-oncology assays.

email: *shubing_wang@merck.com*

12d. A NOVEL ESTIMATION TECHNIQUE FOR A 5-PARAMETER BIVARIATE BETA DISTRIBUTION

Lauren G. Perry*, University of California, Riverside

James M. Flegal, University of California, Riverside

This research proposes a new estimation technique for the five-parameter bivariate beta distributions developed by Arnold and Ng (2011). Since these models lack a closed form density, traditional estimation techniques are unavailable. An estimation method proposed by Arnold and Ng (2011) and referred to as "modified maximum likelihood estimation", uses maximum likelihood estimation on

the marginals to obtain four estimating equations. A fifth equation is obtained through a carefully crafted expectation via method of moments. Unfortunately, z1*z2 appears in the denominator of this expectation, rendering this estimation technique less useful as either z1 or z2 approaches zero. A second estimation method proposed by Crackel and Flegal (2014) avoids this issue, but is highly computationally intensive. This research focuses on novel estimation techniques that are robust to small values of z1 or z2 and are computationally efficient. Simulation studies were run using various parameter values to evaluate the performance of the proposed estimation techniques. Finally, these estimation techniques will be applied to a real-world data set.

email: *lperr003@ucr.edu*

12e. INTERVAL ESTIMATION OF RATIO OF TWO COEFFICIENTS OF VARIATION FOR LOGNORMAL DISTRIBUTIONS

Jun-Mo Nam, National Cancer Institute, National Institutes of Health

Deukwoo Kwon*, University of Miami

Reliability of measurement is fundamental to most studies. The coefficient of variation (CV) can be used as an index of reliability of measurement. The lognormal distribution has been applied to fit data in many fields. We developed approximate interval estimation of the ratio of two CVs for lognormal distributions by using the Wald-Type method, Fieller-Type method, log method, and method of variance estimates recovery (MOVER). Results of simulations show that empirical coverage rates of the above methods are satisfactorily close to a nominal coverage rate for medium sample sizes. For small sample sizes, empirical coverage rates are moderately close to the nominal one when standard deviations are small, however, Fieller-Type method may perform poorly when standard deviations are large, and caution should be used for this case. Among the four methods, MOVER provides the most close coverage rate to a nominal one in average. We present a related test of significance for homogeneity of two CVs using Wald-Type method and provide power and sample size formulae. For a numerical example, we applied the proposed methods to assess the relative reliability of a novel assay method compared to a standard radioimmuno assay in the measurement of estrogen metabolites.

email: *DKwon@med.miami.edu*

12f. EVALUATING R PACKAGES FOR COMPARING TWO CORRELATED C INDICES WITH A RIGHT-CENSORED SURVIVAL OUTCOME

Brian S. Di Pace*, Virginia Commonwealth University

Le Kang, Virginia Commonwealth University

The Concordance (C) index is the probability that a patient with a higher predictive score also has a longer survival time in a randomly drawn pair of subjects. It is important to compare two correlated C indices to determine which score better predicts right-censored survival. An estimator for the variance of the difference in two C indices by Kang et al. does not require resampling and thus is computationally efficient (R package {compareC}). In this simulation study, we generated two scores from a bivariate normal distribution using fixed variances and non-constant means. The covariance structure was defined by the desired correlations. The true variance of the difference was calculated through Monte Carlo Simulations. This variance was estimated using bootstrap resampling and the {compareC} package, and compared to the true variance at varying sample sizes. The empirical type I error rate was calculated for {compareC} and compared to the bootstrap method and existing R functions: cindex.comp, rcorrp.cens and survC1. The results indicate the bootstrap method overestimates the variance, with bias increasing as correlation increases. The {compareC} package has a consistently smaller bias, compared to the bootstrap method, and performs universally well with type I error close to nominal levels.

email: *dipacebs@vcu.edu*

12g. COMPARISON OF TWO CORRELATED ROC CURVES AT A GIVEN SPECIFICITY LEVEL

Leonidas E. Bantis*, University of Texas MD Anderson Cancer Center

Ziding Feng, University of Texas MD Anderson Cancer Center

The receiver operating characteristic (ROC) curve is the most popular statistical tool for evaluating the discriminatory capability of a given continuous biomarker. The need to compare two correlated ROC curves arises when individuals are measured with two biomarkers, which induces paired and thus correlated measurements. Many researchers have focused on comparing two correlated ROC curves in terms of the area under the curve (AUC), which summarizes the overall performance of the marker. However, particular values of specificity may be of interest. We focus on comparing two correlated ROC curves at a given specificity level. We propose parametric approaches, transformations to normality, and nonparametric kernel-based approaches. Our methods can be straightforwardly extended for inference in terms of $ROC^{-1}(t)$. This is of particular interest for comparing the accuracy of two correlated biomarkers at a given sensitivity level. Extensions also involve inference for the AUC and accommodating covariates. We evaluate the robustness of our techniques through simulations and present a real data application involving prostate cancer screening.

email: *leobantis@gmail.com*

12h. LOCALLY RELEVANT SUBGRAPHS ENUMERATION IN TRANSPLANT PATIENT NETWORKS

Wen Wang*, University of Michigan

Mathieu Bray, University of Michigan

Peter Song, University of Michigan

John Kalbfleisch, University of Michigan

In the kidney paired donation (KPD) program, candidate-and-willing-but-incompatible-donor pairs constitute a patient network (PN) under common interest of organ exchanges to achieve mutual benefits from kidney transplants. In PN, altruistic donors (ADs), who are willing to donate and with no associated candidates, are special participants and different from pairs. The PN is formulated as a directed graph. Each vertex represents a pair or an AD; edges denote donations from donors to compatible candidates. Sets of potential transplants are given by subgraphs: disjoint cycles among pairs and chains initiated by ADs. A practical KPD is subject to uncertainties on both edges and vertices due to various reasons (e.g. reneging). To counteract uncertainties, recent research suggests organizing transplants in locally relevant subgraphs, termed as strongly connected subgraphs (SCSs). Within each SCS, cycles and chains cannot be divided into two nonempty parts without common vertices. A key technical task in managing KPD PNs is SCS enumeration. With certain assigned utilities for connectivity, an optimal transplant allocation is disjoint SCC subgraphs maximizing sum of utilities. We develop a breadth-first search based SCS enumeration algorithm. Using simulation study we illustrate that the proposed algorithm is computationally more efficient than the naïve combinatorial algorithm.

email: *wangwen@umich.edu*

12i. STATISTICAL METHODS TO ADDRESS OUTCOME MISCLASSIFICATION IN STUDIES OF ALZHEIMER'S DISEASE

Le Wang*, University of Pennsylvania

Rebecca Hubbard, University of Pennsylvania

Estimates of the relationship between an outcome and an exposure are biased in the presence of imperfect ascertainment of the outcome of interest. Past work has demonstrated that by incorporating the sensitivity and specificity of the imperfect outcome classification, the true association can be recovered and unbiased estimates of the association between disease and risk factors can be obtained. However, methods for misclassified outcomes have seen relatively little use in the context of time to event outcomes. Accounting for outcome misclassification in the context of Alzheimer's disease (AD) is particularly challenging due to the complex relationship between the clinically observable phenotype and the underlying pathology. A recent study found that higher glucose levels may be a risk factor for dementia, using data from the ACT study, a population-based cohort study of aging and dementia. A question remains whether higher glucose levels are also associated with AD. We investigated the association between glucose and age at onset of AD using data from the ACT study, comparing results of discrete-time proportional hazards models with and without adjustment for misclassification of AD.

email: *lwang0217@gmail.com*

12j. THE INFERENCE TREE SYSTEM FOR ACCOUNTABLE ANALYSES

Brian S. Hernandez*, University of Texas Health Science Center, San Antonio

Emmy Burnett, Rice University

Jonathan A. Gelfond, University of Texas Health Science Center, San Antonio

The American Statistical Association in currently updating its Ethical Guidelines for Statistical Practice. Accountability and reproducibility are being strongly considered as new additions to these guidelines. Hence, these principles are now taken to be fundamental to the practice of statistics, but the definitions and the tools for ensuring reproducibility and accountability are not standardized. In general, accountability is the ability to take responsibility and give a verifiable account of what was done. What is less appreciated is that accountability can be rigorously defined using the terms of computer science. This implies a need for new computing systems and environments that satisfy such definitions. A statistical analysis can be viewed as a directed acyclic graph (DAG) in which data and computer programs are nodes. The input nodes would be the data and the output nodes would be the resultant computations, which could be input nodes for dependent actions. We have developed a system called Inference Tree in R that is built on the principle of accountable units. An accountable unit is a data file (statistic, table or graphic) that can be associated with a provenance, meaning how it was created, when it was created and who created it. It works by integrating a version control system, cryptographic hashes, and a dependency tracking database.

email: *hernandezbs@uthscsa.edu*

12k. HIV INCIDENCE ESTIMATION FROM A CROSS-SECTIONAL SURVEY: AN APPROACH TO CALIBRATE A BIOMARKER AND DERIVE THE MLE OF THE INCIDENCE

Severin Guy Mahiane*, Avenir Health

We present two statistical methods to study the distribution of a biomarker and derive a formula for estimating HIV incidence from a cross-sectional survey. Both methods allow handling interval censored data and basically consist of using a generalized mixture model to

describe the trajectory of the biomarker as a function of time since infection. The first uses data from all followed-up individuals and allows incidence estimation in the cohort, whereas the second only uses data from seroconverters. We illustrate our methods using repeated measures of the IgG capture BED enzyme immunoassay. Estimates of calibration parameters, that is, mean window period, mean recency period, sensitivity, and specificities obtained from both models are comparable. The formula derived for incidence estimation gives the maximum likelihood estimate of incidence which, for a given window period, depends only on sensitivity and specificity. The optimal choice of the window period is explored. Numerical simulations suggest that data from seroconverters can provide reasonable estimates of the calibration parameters.

email: *GMahiane@avenirhealth.org*

12l. A SIMPLE DENSITY-BASED EMPIRICAL LIKELIHOOD RATIO TEST FOR INDEPENDENCE

Albert Vexler, University of Buffalo, The State University of New York

Wan-Min Tsai*, PPD and University of Buffalo, The State University of New York

Alan Hutson, University of Buffalo, The State University of New York

We develop a novel nonparametric likelihood ratio type test for independence between two random variables using a technique that is free of the common constraints of defining a given set of specific dependence structures. Our methodology revolves around an exact density-based empirical likelihood ratio test statistic that approximates in a distribution-free fashion the corresponding most powerful parametric likelihood ratio test. We demonstrate that the proposed test is very powerful relative to detecting general structures of dependence between two random variables, including non-linear and/or random-effect type of dependence structures. An extensive Monte Carlo study confirms that the proposed test is superior to the classical nonparametric procedures across a variety of settings. In some instances, the power gain is dramatic. The real-world applicability of the proposed test is illustrated using datasets from a study of the role of vitamin D in the pathogenesis of thromboembolism and a study of biomarkers associated with myocardial infarction.

email: *wanmin1027@gmail.com*

13. POSTERS: REPEATED MEASURES

13a. CAREFUL CONSIDERATION OF TIME-VARYING EXPOSURES WITH POSSIBLE REPEATED EVENTS

Andrew D. Althouse*, University of Pittsburgh

Patients with end-stage heart failure may be implanted with a left ventricular assist device (LVAD) to prolong survival and possibly bridge patients to heart transplant. However, these patients have a very high risk for subsequent adverse events (AE's) such as infection, bleeding, and device malfunction. Our primary aim was to evaluate the relationship between treatment non-compliance (NC) after device implant and risk of subsequent AE's. Selecting the most appropriate analysis for this question is not a straightforward task. The exposure (NC) occurs at different times for each patient, and its effects on AE risk may be rather transient. Furthermore, patients may have multiple occurrences of each AE. Finally, there is the question of what to do with the initial period of "compliant" time for patients who are eventually NC. This presentation will discuss several analytic options and the merits of each. Cox proportional-hazards models and negative binomial regression will both be considered, and within those modeling choices, particular attention will be devoted to the following questions: How should time at risk be assigned to patient-time of Compliance vs. Non-Compliance? How can we capture the full burden of Non-Compliance by accounting for multiple events?

email: *althousead@upmc.edu*

13b. MAXIMUM-LIKELIHOOD BASED ANALYSIS OF KIDNEY TRANSPLANT CENTER REPORT CARDS

Shaun D. Bender*, University of Pennsylvania

Peter P. Reese, University of Pennsylvania

Victoria Gamerman, Boehringer Ingelheim Pharmaceuticals Inc.

Justine Shults, University of Pennsylvania

Report cards for kidney transplant centers were developed with the goal of improving the experience of transplant patients, by motivating centers who were flagged for poor performance to make changes to improve their quality of care. However, being flagged in a bi-annual public report could have unintended negative consequences. For example, a negative report could discourage a center from treating high-risk patients, which could result in a reduction in the total number of transplants performed, as well as the number of transplants with live donors. We present a maximum likelihood based approach for analysis of these transplant center outcomes. The proposed method accounts for the serial correlation and over-dispersion, and can be applied to data that are unequally spaced in time.

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

13c. A FLEXIBLE APPROACH FOR ANALYZING LONGITUDINAL CLUSTERED DATA: A GENERALIZATION OF THE DIFFERENCE-IN-DIFFERENCE (DD) APPROACH

Jason A. Lee*, University of Florida
W. Bruce Vogel, University of Florida
Martin P. Wegman, University of Florida
Keith E. Muller, University of Florida

The difference-in-difference (DD) approach, often used in evaluations of policy implementation and quasi-experimental designs, compares outcome variable differences between two groups at two time points to isolate and test the presence of a treatment effect, assumed to be the deviation from the baseline difference. We propose a mathematical generalization of the DD approach allowing the simultaneous modeling of more than two time points while accounting for differing slopes between groups over time. This is accomplished by placing baseline response values into a general linear mixed model as a covariate and assuming an unstructured covariance matrix over time. Our approach increases overall model power by borrowing strength over time. We allow the baseline difference between groups on the treatment effect to be any value, unlike the DD approach, which is restricted to 1.0 due to the parallel trend assumption. The generalization of the DD approach provides a robust framework which facilitates treatment effect estimation in data, particularly where model convergence is inhibited by the presence of cluster imbalance and missingness. Expanded use of our generalization could provide a more complete view of policy effect patterns and better the quality of policy evaluation results.

email: *jasandlee1@gmail.com*

13d. METHODS FOR EVALUATING RESPONDENT ATTRITION IN ONLINE SURVEY DATA

Camille J. Hochheimer*, Virginia Commonwealth University
Roy T. Sabo, Virginia Commonwealth University
Alexander Krist, Virginia Commonwealth University
Steven H. Woolf, Virginia Commonwealth University
Teresa Day, Virginia Commonwealth University

With health services research expanding to online formats, patients have more freedom and control to participate in surveys with the ability to skip questions or quit before completion. Using web-based paradata, researchers can capture not only a patient's responses but also which questions they skip and if/when they drop out. Despite a call for specific methods to evaluate attrition in online survey data, there is almost no published methodology on the subject. Using an informed decision-making (IDM) module that our research team fielded through an interactive online patient portal as a motivating example, this study explores ways to quantify attrition, identify significant attrition patterns, compare attrition rates between cohorts using discrete time survival analysis, and determine covariates associated with dropout. We intend for this study to serve as a starting point for further exploration into attrition analysis.

email: *hochheimercj@vcu.edu*

13e. PREDICTING SLEEP STAGES VIA GAUSSIAN PROCESSES

Xu Gao*, University of California, Irvine
Hernando Ombao, University of California, Irvine
Babak Shahbaba, University of California, Irvine

The goal of this paper is to develop a model for predicting sleep stage, which is a binary time series, using a novel two-stage classification method. To this end, we model the time series using mixed effects latent processes, where the fixed term captures the effect of covariates (heart rate/body temperature) and the random effect is modeled using a Gaussian process on time domain. This way, the proposed method provides high prediction accuracy. Moreover, approaches on model selection are also developed. Results suggest the benefit of using the proposed method over logistic regression in terms of higher and more stable prediction accuracy. Test results also show that the proposed method is promising when missing values occur. Laplace approximation, golden section search and successive parabolic interpolation are also utilized to control the computational complexity.

email: *xgao2@uci.edu*

13f. AN EXTENSION OF AUTOREGRESSIVE AND CROSS-LAGGED MODELS TO MODELING CORRELATED BIVARIATE NON-COMMENSURATE OUTCOMES

Fei He*, Indiana University
Armando Teixeira-Pinto, University of Sydney
Jaroslaw Harezlak, Indiana University School of Public Health

WITHDRAWN

Autoregressive and cross-lagged models have been widely used to understand the relationship between multiple commensurate outcomes in social and behavioral sciences, but not much work has been done in modeling multiple correlated non-commensurate outcomes simultaneously. We developed a likelihood-based methodology combining ordinal autoregressive and cross-lagged models with a shared subject-specific random effect in the mixed model framework to model two correlated longitudinal non-commensurate outcomes. The estimates of the cross-lagged and the autoregressive effects from our model were shown to be consistent and were more efficient than the estimates from the univariate generalized linear models. Inclusion of the subject-specific random effects in

the proposed model accounted for between-subject variability arising from the omitted subject-level predictors. Our model is not restricted to the case with equal number of events per subject and it can be easily extended to three or more non-commensurate outcomes. We applied our model to an ecological momentary assessment study with complex dependence and sampling data structures. Specifically, we studied the dependence between the condom use and sexual satisfaction based on the data reported in a longitudinal study of sexually transmitted infections. We found negative cross-lagged effect between these two outcomes and positive autoregressive effect within each outcome.

email: *hefei@iupui.edu*

13g. OBSERVATIONS OR EVENTS PER VARIABLE IN LONGITUDINAL MODELS

Abigail R. Smith, Arbor Research Collaborative for Health

Jarcy Zee*, Arbor Research Collaborative for Health

It is well known that model overfitting can lead to spurious associations. Rules of thumb are often used to guide the number of observations or events per variable needed to fit multivariable regression models with continuous or dichotomous outcomes, respectively. However, current guidelines only apply to studies with independent observations, and no such guidelines have been established in the setting of clustered or longitudinal data analysis. At the extremes, either the total number of observations/events or the total number of clusters could be used with current guidelines to determine the number of parameters that can be estimated in a model. However, a more accurate measure likely lies somewhere in between. We propose to use the design effect to calculate an effective number of observations or events, and then use this to determine the maximum number of covariates that can be included without overfitting the longitudinal model. Through a simulation study, we tested the performance of our proposed method under different scenarios, including continuous and binary outcomes and predictors, as well as correlated predictors and varying levels of intraclass correlation. We compared these results to those using both the total number of observations/events as well as the total number of clusters.

email: *Jarcy.Zee@arborresearch.org*

13h. MAXIMUM LIKELIHOOD BASED ANALYSIS OF EQUALLY SPACED LONGITUDINAL COUNT DATA WITH SPECIFIED MARGINAL MEANS, FIRST-ORDER ANTEDEPENDENCE, AND LINEAR CONDITIONAL EXPECTATIONS

Victoria Gamerman*, Boehringer-Ingelheim Pharmacueticals, Inc. and University of Pennsylvania

Matthew Guerra, U.S. Food and Drug Administration

Justine Shults, University of Pennsylvania

This work implements a maximum likelihood based approach that is appropriate for equally spaced longitudinal count data with over-dispersion, so that the variance of the outcome variable is larger than expected for the assumed Poisson distribution. We implement the proposed method in the analysis of two data sets and make comparisons with the semi-parametric generalized estimating equations approach. The simulations demonstrate that the proposed method has better small sample efficiency than the GEE approach that incorrectly ignores the over-dispersion. We also provide user-written code in R that can be used to recreate the analysis results.

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

13i. PARSIMONIOUS REGRESSION MODELS FOR ASSOCIATIONS OF ACCELEROMETRY-DERIVED FEATURES OF WALKING AND PERFORMANCE MEASURES IN THE ELDERLY POPULATION

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 Public Health
Objective methods for measuring physical activity data rely heavily on the ever more used wearable accelerometers. Our work focuses on walking characteristics derived from the high-density accelerometry data collected in both free-living and in-the-lab conditions. We propose robust methods of characterizing both micro- and macro-scale walking features. At micro-scale, we focus on cadence (expressed in steps-per-second) and walking-specific acceleration level (expressed in g-units). For macro-scale analysis we focus on daily walking time and total acceleration produced. Proposed feature extraction methodology is based on the authors' prior work on the identification of periods of sustained harmonic walking. Data for the analysis were collected on thirty-eight community-dwelling elderly individuals, 17 males and 21 females, by hip-worn, tri-axial ActiGraph accelerometers. We build parsimonious regression models for the association between the proposed walking characteristics and outcomes of standardized performance tests including measures of cognitive executive functioning, fatigability, physical functionality, mobility function and self-reported caloric expenditure. We show that the application of the statistical feature extraction methods to the high-density activity data collected in the free-living settings provides additional information

beyond the commonly used summary measures and widely used in-the-lab experiments.

email: *jurbane2@jhu.edu*

13j. EMPIRICAL BAYES SHRINKAGE ESTIMATORS FOR SUMMARY STATISTICS OF NON-STATIONARY TIME SERIES

Amanda F. Mejia*, Johns Hopkins Bloomberg School of Public Health

Ciprian Crainiceanu, Johns Hopkins Bloomberg School of Public Health

Martin Lindquist, Johns Hopkins Bloomberg School of Public Health

Shrinkage estimators have been shown to outperform subject-level observations in a variety of fields, including neuroimaging (Shou et al. 2014, Mejia et al. 2015). For example, empirical Bayes shrinkage estimators of functional connectivity measures derived from fMRI data have been shown improve reliability of subject-level connectivity estimates (Shou et al. 2014) and parcellations (Mejia et al. 2015). In Mejia et al. (2015), a shrinkage method was proposed for single-session fMRI data based on using "pseudo scan-rescan data" to estimate the within-subject variance, which was adjusted using a factor estimated empirically. Here, we revisit the estimation of within-subject variance for a summary statistic derived from a non-stationary time series. In particular, we present a different measurement error model and show that the raw estimator contains two sources of within-subject variability: sampling variance and within-subject signal variance. We propose a method to estimate both sources of variability, which allows for estimation of the total within-subject variability for the purpose of shrinkage. We demonstrate the proposed method on several test-retest fMRI datasets and show that it results in improved reliability over the previously proposed methods. Shou, Haochang, et al. "Shrinkage prediction of seed-voxel brain connectivity using resting state fMRI." NeuroImage 102 (2014): 938-944. Mejia, Amanda F., et al. "Improving reliability of subject-level resting-state fMRI parcellation with shrinkage estimators." NeuroImage 112 (2015): 14-29.

email: **amejia@jhsph.edu**

13k. MODIFIED QUASI-LIKELIHOOD CRITERION FOR GENERALIZED ESTIMATING EQUATIONS

Chelsea B Deroche*, University of Missouri

The use of generalized estimating equations (GEEs) in the public health and medical fields are important tools for dealing with the within group correlation for longitudinal, clustered, or panel data. Generalized estimating equations are an extension to the generalized linear model specifically modified to address the within group correlation. The current quasi-likelihood information criterion (QIC) for selecting the working correlation structure is not efficient: it tends to favor the independent structure which assumes there is no within group correlation. I propose a modified QIC that outperforms the current QIC in that this criterion favors the correct structure a large majority of the time. This modified QIC not only takes into account the number of parameters in the model, but also accounts for the number of correlation estimates. Model building in generalized estimating equations and the simulation results of the modified QIC will be presented.

email: *derochec@health.missouri.edu*

13l. AN EMPIRICAL APPROACH TO DETERMINE A THRESHOLD FOR DECLARING THE PRESENCE OF OVERDISPERSION IN COUNT DATA

Elizabeth Payne*, Medical University of South Carolina

Overdispersion is a problem commonly encountered in count data which can lead to invalid inference if left unaddressed. One of the most commonly used estimators of dispersion in the literature is the ratio of the Pearson χ^2 statistic to its corresponding degrees of freedom. Decision about whether data are overdispersed is typically made by checking whether this ratio is bigger than one. It is not clear how far a deviation from one of this ratio leads to wrong inference. That is, there is currently no fixed threshold for declaring the necessity of statistical intervention for dealing with overdispersion. In this paper, we consider simulated datasets containing varying magnitudes of outlier dependent overdispersion in order to empirically determine an appropriate threshold value of the ratio of Pearson χ^2 to degrees of freedom for determining the level of overdispersion that leads to wrong statistical inference if the analysis ignores overdispersion. We consider both Poisson and negative binomial regression and generalized linear mixed model approaches and utilize Type 1 and Type 2 errors and confidence interval coverage probabilities to measure the effect of different levels of overdispersion on inference.

email: *payneeh@musc.edu*

14. POSTERS: SPECIAL TOPIC

14a. INVITED POSTER: THE INTERNATIONAL BIOMETRIC SOCIETY

Elizabeth Thompson*, University of Washington, IBS President

The International Biometric Society (IBS) is the international society that promotes the development and application of statistical and mathematical theory and methods in the biosciences, includ-

ing agriculture, biomedical science and public health, ecology, environmental sciences, forestry, and allied disciplines. The Society publishes two journals, Biometrics and JABES (the Journal of Agricultural, Biological, and Environmental Statistics). The IBC meetings, every two years, bring together statistical scientists from a broad variety of countries and a diversity of disciplines: we meet in Victoria BC in 2016, and in Barcelona, Spain, in 2018. The IBS consists of 34 regions, including 19 regions that include members from developing countries, as defined by the World Bank. The East North American Region (ENAR) is, by count of members, the largest region of the IBS. While many ENAR members participate in the activities of the IBS, many others do not. The goal of this poster is to introduce ENAR members to recent IBS activities and developments, and to encourage broader participation, especially from younger members.

email: *eathomp@u.washington.edu*

15. STATISTICAL ADVANCES IN FUNCTIONAL AND SINGLE CELL GENOMICS

TOWARDS A GLOBAL GENE REGULATORY NETWORK

Wing Hung Wong*, Stanford University

Yong Wang, Stanford University

Rui Jiang, Tsinghua University

With the development of new methods such as ATAC-seq, it is now easy to map the locations of open chromosomal regions from a small number of cells. Thus regulome profiling has joined expression profiling as the two pillars of high throughput functional genomics. In this talk I will review progress towards building a global model for gene regulatory relations based on these two types of information.

email: *whwong@stanford.edu*

A SPECTRAL APPROACH FOR THE INTEGRATION OF FUNCTIONAL GENOMICS ANNOTATIONS FOR BOTH CODING AND NONCODING SEQUENCE VARIANTS

Iuliana Ionita-Laza*, Columbia University

Kenneth McCallum, Columbia University

Bin Xu, Columbia University

Joseph Buxbaum, Mount Sinai School of Medicine

Over the past few years, substantial effort has been put into the functional annotation of variation in human genome sequence. Such annotations can play a critical role in identifying putatively causal variants among the abundant natural variation that occurs at a locus of interest. The main challenges in using these various annotations include their large numbers, and their diversity. I will discuss an unsupervised approach to derive an integrative score of these diverse annotations. I will show that the resulting meta-score has good discriminatory ability using disease associated and putatively benign variants from published studies (for both Mendelian and complex diseases), and is more strongly associated with the disease association status of such variants compared with the recently proposed CADD score. Furthermore, I will show how the meta-score is particularly useful in prioritizing likely causal variants in a region of interest when it is combined with sequencing data in the framework of a hierarchical model.

email: *ii2135@cumc.columbia.edu*

STATISTICAL MODELING OF DROPOUT EVENTS IN SINGLE-CELL RNA SEQUENCING DATA

Mingyao Li*, University of Pennsylvania

Cheng Jia, University of Pennsylvania

Yuchao Jiang, University of Pennsylvania

Nancy Zhang, University of Pennsylvania

Single-cell RNA-seq has made it possible to examine transcriptomic variations at single cell levels. Examining transcriptomic variations from individual cells will provide a higher resolution of cellular differences and a better understanding of the function of an individual cell in its microenvironment. However, careful analysis is required in order to answer biological questions because a low starting amount of mRNA often makes it more likely that a transcript will be missed during library preparation and consequently not detected during sequencing, and this will lead to the so-called "dropout" events. In this talk, I will present a novel statistical framework that explicitly models those "dropout" events. It is based on the key observation that the probability of being a "dropout" depends on the true expression of a gene in that a lowly expressed gene is more likely to be a dropout. Using this framework, we are able to infer the true expression of a gene, detect allele specific gene expression, and differential gene expression between different conditions. I will illustrate our methods using both simulated data as well as real single-cell RNA-seq data. This is based on joint work with Nancy Zhang.

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

A DIRICHLET PROCESS MIXTURE MODEL APPROACH TO IDENTIFY GENES SHOWING DIFFERENTIAL DYNAMICS IN SINGLE-CELL RNA-Seq DATA

Keegan Korthauer, Dana-Farber Cancer Institute

Rhonda Bacher, University of Wisconsin, Madison
Jeea Choi, University of Wisconsin, Madison
Li-Fang Chu, Morgridge Institute for Research
James A. Thomson, Morgridge Institute for Research
Ron Stewart, Morgridge Institute for Research
Christina Kendziorski*, University of Wisconsin, Madison

Identifying genes with average expression that varies across two or more biological conditions, so-called differentially expressed genes, has proven useful in a multitude of bulk RNA-seq studies. However, because cell-to-cell heterogeneity is masked in a bulk experiment, many important types of differences go unobserved. A gene that appears to be equivalently expressed in a bulk experiment, for example, may actually manifest different proportions of cells in sub-populations across conditions. Alternatively, there may be a distinct sub-population that presents in a given condition. Single-cell RNA-seq provides the opportunity to identify cell sub-populations within a condition and to characterize genes with differential dynamics (DD) across conditions, but the statistical methods available for doing so are greatly limited, largely because they do not fully accommodate the cell heterogeneity that is prevalent in single-cell data. To address this, we developed a Bayesian modeling framework to facilitate the characterization of expression within a biological condition, and to identify DD genes across conditions in a scRNA-seq experiment. Simulation studies suggest that the approach provides improved power and precision for identifying DD genes. Additional advantages are demonstrated in a case study of human embryonic stem cells.

email: *kendzior@biostat.wisc.edu*

16. STATISTICAL CONSIDERATIONS AND CHALLENGES IN EVALUATING VACCINE EFFICACY

EVALUATING EBOLA VACCINE EFFICACY UNDER OUTBREAK CONDITIONS USING A RING VACCINATION TRIAL DESIGN

Natalie E. Dean*, University of Florida
Ira M. Longini, University of Florida
M. Elizabeth Halloran, Fred Hutchinson Cancer Research Center and University of Washington

Evaluating the efficacy of a vaccine during a public health emergency presents important challenges. In this talk we describe, in detail, the novel ring vaccination trial design that went into the field in Guinea in March 2015. Modeled after the strategy used to eradicate smallpox, rings are comprised of contacts and contacts of contacts of Ebola-infected index cases. Rings are then cluster-randomized to receive vaccine immediately or in a delayed fashion. Comparison of Ebola incidence in the early vs. delayed rings forms the basis for the estimation of vaccine efficacy. An interim analysis of the trial suggests 100% vaccine efficacy 10 or more days after randomization (95% CI: 74.7-100.0). We describe the analysis of the trial and discuss key operational and statistical issues. We explore how the ring vaccination trial design may be an effective model for evaluating vaccine efficacy during future infectious disease outbreaks.

email: *nataliedean@ufl.edu*

INFERENCE ABOUT HERD IMMUNITY IN OBSERVATIONAL VACCINE STUDIES

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

Quantifying indirect (or spillover) effects of vaccination, i.e., herd immunity, is important from a public health perspective. In the nomenclature of causal inference, such indirect effects are present when there is interference between individuals, i.e., when the treatment (vaccination) of one individual affects the outcome of another individual. In this talk we will consider recently developed methods for inference about treatment effects in the presence of interference. The methods will be illustrated with data from an individually-randomized, placebo controlled trial of cholera vaccination in 122,000 individuals in Matlab, Bangladesh which indicates significant evidence of herd immunity.

email: *mhudgens@bios.unc.edu*

SIEVE ANALYSIS USING THE NUMBER OF INFECTING PATHOGENS

Dean A. Follmann*, National Institute of Allergy and Infectious Diseases, National Institutes of Health
Ching-Yu Huang, Johns Hopkins University

Assessment of vaccine efficacy as a function of the similarity of the infecting pathogen to the vaccine is an important scientific goal. Characterization of pathogens (e.g., different strains of HIV or malaria) for which vaccine efficacy is low can increase understanding of the vaccine's mechanism of action and offer targets for vaccine improvement. Traditional sieve analysis estimates differential vaccine efficacy using the time to infection by a type of competing risks survival analysis. Each infected subject has a single pathogen identified, e.g. by consensus at each amino acid location, and the similarity between this pathogen and the vaccine inserts quantified e.g. exact match or number of mismatched amino acids. With new technology we can now obtain the actual count of genetically distinct pathogens that infect an individual. This talk introduces new methods for sieve analysis that exploit this count information. Let A be the number of distinct pathogens. We construct an A-dimensional multivariate process with individual process "a" jumping from 0

to the number of type "a" infecting pathogens at the common time of infection, for $a=1,...,A$. The process is based on a log-linear regression model for the per-exposure expected number of type "a" infecting pathogens coupled with a proportional hazards model for the common time to infection. The log-linear regression model allows parasimonious quantification of the "distance" between pathogen "a" and the vaccine insert, e.g. perfect match/mismatch at a sub-region, or total number of mismatches. We derive estimating equations for the log-linear regression parameters from which we estimate differential vaccine efficacy. The new method is compared to traditional sieve methods by simulation and applied to a malaria vaccine study.

email: *dean.follmann@nih.gov*

17. RECENT ADVANCES IN SUBGROUP IDENTIFICATION FOR CLINICAL TRIAL REGULATORY SCIENCE

A BAYESIAN CREDIBLE SUBGROUPS APPROACH TO IDENTIFYING PATIENT SUBGROUPS WITH POSITIVE TREATMENT EFFECTS

Bradley P. Carlin*, University of Minnesota

Patrick M. Schnell, University of Minnesota
Qi Tang, AbbVie, Inc.

Walter W. Offen, AbbVie, Inc.

Many new experimental treatments benefit only a subset of the population. Identifying the baseline covariate profiles of patients who benefit from such a treatment, rather than determining whether or not the treatment has a population-level effect, can substantially lessen the risk in undertaking a clinical trial and expose fewer patients to treatments that do not benefit them. The standard analyses for identifying patient subgroups that benefit from an experimental treatment either make separate marginal inferences on each individual, which raises multiplicity issues, or focus inappropriately on the presence or absence of treatment-covariate interactions. We propose a Bayesian "credible subgroups" method to identify two bounding subgroups for the benefiting subgroup: one for which it is likely that all members simultaneously have a treatment effect exceeding a specified threshold, and another for which it is likely that no members do. We examine frequentist properties of our method via simulation, and illustrate the approach using data from an Alzheimer's disease treatment trial. Time permitting, we discuss extension to settings in which there are multiple test treatments and multiple control groups, as well as those in which multiple measures of efficacy or safety are of interest.

email: *brad@biostat.umn.edu*

DETECTION OF PREDICTIVE BIOMARKERS ACCOUNTING FOR SAMPLE HETEROGENEITY

Jianhua Hu*, University of Texas MD Anderson Cancer Center

Weining Shen, University of California, Irvine

Jing Ning, University of Texas MD Anderson Cancer Center

Zideng Feng, University of Texas MD Anderson Cancer Center

Cancer heterogeneity is well recognized in biomedical studies. Such heterogeneity can help or hinder important discoveries depending on whether or not it is properly accounted for in data analysis. In biomarker validation studies involving targeted biomarkers, we intend to identify biomarkers and/or clinical variables that characterize patient subgroups associated with specific clinical outcomes, which can potentially improve cancer early detection, diagnosis, and prognosis. Specifically, we propose a structured logistic-mixture model and develop a formal statistical testing procedure to identify important factors that determine patient subgroups and are directly associated with binary clinical outcomes.

email: *jhu@mdanderson.org*

EVALUATING THE IMPACT OF TREATING THE OPTIMAL SUBGROUP

Alexander R. Luedtke, University of California, Berkeley

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

Suppose we have a binary treatment used to influence an outcome. Given data from an observational or controlled study, we wish to determine whether or not there exists some subset of observed covariates in which the treatment is more effective than standard practice. Furthermore, we wish to quantify the improvement in population mean outcome that will be seen if this subgroup receives treatment and the rest of the population remains untreated. This problem is surprisingly challenging given how often it is an (at least implicit) study objective. Blindly applying standard techniques fails to yield any asymptotically useful results, while using existing techniques to confront the non-regularity does not appear to help. We will describe an approach to estimate the impact of treating the subgroup which benefits from treatment that is valid in a nonparametric model. This approach requires that we consider stochastic interventions, where the probability of an individual receiving treatment increases according to their probability of belonging to the subgroup which benfits from treatment. We will argue that this stochastic approximation may be more interesting than the deterministic alternative. This is joint work with Alex Luedtke.

e-mail: *laan@berkeley.edu*

18. CENS INVITED SESSION: WHAT I KNOW NOW: ADVICE ON MAXIMIZING GRADUATE SCHOOL AND EARLY CAREER EXPERIENCE

WHEN DO WE BECOME DINOSAURS: LIFE AFTER GRAD SCHOOL

Janet Wittes*, Statistics Collaborative

Each of us enters graduate school as a sort of tabula rasa (or, partially rasa). Our professors fed us with theory and methods which we digested as best we were able. In fact, as graduate students, our job was to learn. We leave school with a diploma in hand, a mortar board on our head, and newly gained knowledge in our minds. The immediate Incentives for learning we had become used to in graduate school decrease precipitously when we enter our post-graduate school lives. At first we can successfully coast on our previous achievements. As the years pass, however what we have learned often becomes dated; if we fail to keep up with the literature, we can no longer contribute meaningfully to discussion. As we age, our physical ability to concentrate becomes more difficult and it is harder to learn new skills. If we allow ourselves to succumb to the fiction that our current skills are sufficient to our continued ability to contribute effectively as statisticians, we become professional dinosaurs. This talk addresses the need for continued learning and discusses strategies than are useful for preventing intellectual ossification.

email: *janet@statcollab.com*

STRATEGIC PLANNING AND MANAGEMENT OF YOUR ACADEMIC CAREER IN BIOSTATISTICS

Richard John Cook*, University of Waterloo

There are several multi-faceted dimensions to a typical academic position including research, graduate student supervision, teaching, and professional leadership. This talk will highlight these different dimensions and point to ways that useful experience and progress can gained during graduate school and early in an academic position. Being aware of these different dimensions and their relative importance to any particular position, as well as careful monitoring of progress, can help put individuals in the best position for success and promotion. The engagement of mentors, the need for critical self-evaluation, and the importance of embracing opportunities and challenges will be emphasized.

email: *rjcook@uwaterloo.ca*

STATISTICAL THEORY, POLICY FACT: PREPARING FOR THE ROLE OF A GOVERNMENT STATISTICIAN

Steven Hoberman*, U.S. Food and Drug Administration

Statistics, with its overarching goal of quantifying variation and wresting meaning from uncertainty, has a philosophical as well as mathematical component. Government policies, with their aim of furthering stability and the common good, have a philosophical framework as well. The government statistician, well versed in both the techniques and conceptual foundations of their field, must bridge these two worlds. In graduate school, statisticians learn basic principles and theorems as well as more advanced and detailed statistical techniques. In government work, it is the strong command of these basic principles that gives one the flexibility to translate statistical theory into a real world government policy. Issues such as the appropriateness of non-parametric versus parametric techniques, and the design of simulations arise frequently. Familiarity with SAS and R is also important. This talk discusses skills that are useful for this career path.

email: *Steven.Hoberman@fda.hhs.gov*

19. PRECISION MEDICINE: STATISTICAL CHALLENGES AND OPPORTUNITIES

MACHINE LEARNING AND PRECISION MEDICINE

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

Precision medicine seeks to leverage patient heterogeneity to provide reproducible, optimized treatment regimens. In this talk, we will discuss recent developments in machine learning which have great potential to advance the precision medicine quest. These new methods can help clarify causes of treatment heterogeneity and can discover treatment rules involving complex and multi-stage treatment options, including options ranging over a continuum such as dose level or timing of treatment. We will also present several illustrative clinical examples.

email: *kosorok@unc.edu*

ADAPTIVE TREATMENT ASSIGNMENT: GETTING PERSONAL IN ONCOLOGY

Peter F. Thall*, University of Texas MD Anderson Cancer Center

Physicians have been practicing personalized medicine for thousands of years, using the general paradigm “Observe, act adaptively, repeat.” Today, this may involve within-patient re-randomization, backward induction, machine learning, and design pre-validation by computer simulation. In this talk, I will discuss some clinical trial designs and also my practical experiences with oncologists to design trials that evaluate multi-stage dynamic treatment regimes, based on my work as a biostatistician at M.D. Anderson Cancer Center. I will explain why optimizing multi-stage adaptive therapies may be counter-intuitive, and the difference between a given theoretical

ideal and actual trial conduct. Topics will include a new two-stage dose-finding method that is adaptive both between and within patients, surprisingly reliable bias correction based on a Bayesian nonparametric regression model, a SMART trial in prostate cancer, and recent results from a SMART trial of targeted agents for metastatic renal cancer.

email: *peterthall6775@gmail.com*

IDENTIFYING BIOSIGNATURES FOR PLACEBO RESPONSE USING HIGH DIMENSIONAL FUNCTIONAL DATA

Thaddeus Tarpey*, New York University

Eva Petkova, New York University

Todd Ogden, Columbia University

Jie Vera Tian, Wright State University

One of the difficutlies in determining optimal treatments for various mental illnesses, such as depression, is high placebo response rates. Modeling placebo response when individuals are taking an active treatment (e.g. a pill) continues to be a major statistical challenge. Consequently, the problem of finding biomarkers for placebo response as well as specific drug respons is a critical issue for advancing precision medicine in mental health. The availability of modern high dimensional baseline modalities such as brain imaging data has reinvigorated the goal of finding these potential biomarkers. In this talk, we explore the use of high dimensional functional baseline covariates to help distinguish specific drug response from placebo response using data from a depression trial. In particular, we use a regression model with a functional (longitudinal) outcome and several functional baseline predictors from a placebo-controlled depression study to determine what aspects of the functional covariates are predictive of placebo response and specific drug response.

email: *thaddeus.tarpey@wright.edu*

OPTIMIZING THE PERSONALIZED TIMING FOR TREATMENT INITIATION WITH RANDOM DECISION POINTS

Lu Wang*, University of Michigan

Yebin Tao, University of Michigan

Stepwise intensification of treatment is often necessary for chronic diseases with progressive conditions. An important but challenging problem is to find the optimal personalized timing to initiate a treatment for the next stage of disease condition. In this talk, we consider estimating the optimal dynamic treatment regimes (DTRs) to determine a personalized timing for treatment initiation given a patient's specific characteristics. We aim to identify the optimal DTR amongst a set of regimes predefined by key biomarkers idicating disease severity, which are monitored continuously during a follow-up period. 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 initiation based on patients' biomarker and treatment history. Under each candidate DTR, we employ a flexible survival model with splines of time-varying covariates to estimate the patient-specific probability of adherence to the regime. With the estimated probability, we construct an inverse probability weighted estimator for the counterfactual mean utility (predefined criteria) to assess the DTR. We conduct simulations to demonstrate the performance of our method and further illustrate the application process with an example of insulin therapy initiation among type 2 diabetic patients.

email: *luwang@umich.edu*

20. INNOVATIVE STATISTICAL METHODOLOGY FOR MODELING GROWTH IN FETUSES, CHILDREN, AND ADOLESCENTS

PREDICTING POOR PREGNANCY OUTCOMES FROM MULTIVARIATE ULTRASOUND FETAL GROWTH DATA

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

Developing predictors of poor pregnancy outcomes such as small-for-gestational age or preterm birth is important for monitoring pregnant women. We will begin by presenting simple two-stage estimation procedures that approximates a full maximum-likelihood approach for predicting a binary event from multivariate longitudinal growth data (Albert, Statistics in Medicine, 2012). Subsequently, we will present a class of joint models for multivariate growth curve data and a binary event that accommodates a flexible skewed error distribution for the ultrasound measurements and an asymmetric link function relating the longitudinal to the binary process (Kim and Albert, submitted for publication). Finally, we will present a tree-based approach for identifying subgroups of women who have an enhanced predictive accuracy for predicting a binary event from longitudinal fetal growth data (Foster, et al. submitted for publication). These new statistical methodologies will be illustrated using fetal growth data.

email: *msuchard@ucla.edu*

MODELING CHILDHOOD GROWTH DATA WITH HISTORICAL FUNCTIONAL REGRESSION AND LANDMARKING

Jonathan E. Gellar*, Mathematica Policy Research

Lei Huang, Johns Hopkins Bloomberg School of Public Health

Luo Xiao, North Carolina State University

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

We present a novel approach for modeling childhood growth curves using functional regression techniques. The fundamental goal is to model longitudinal growth data conditional on all previous values of the outcome and time-varying predictors, via smooth functional terms. Potential outcomes in these models can be simple scalar (at a fixed time point), longitudinal (such as the growth curve itself), or survival (such as time to growth stunting). We also extend these methods to the dynamic prediction setting by incorporating landmarking techniques. This approach allows us to address the goal of predicting future trajectories or events based on all information available up to a particular landmark time. Methods are inspired by and applied to a dataset of growth curves from young children in rural Peru, a population with a high degree of growth stunting.

email: *JGellar@mathematica-mpr.com*

SITAR - A SHAPE INVARIANT MODEL FOR HUMAN GROWTH IN INFANCY AND PUBERTY

Tim J. Cole*, University College London Institute of Child Health

Growth reflects size changing over time. However the time scale is not invariant as it varies between individuals according to their underlying developmental age. Thus growth modelling should allow for variability on both the measurement and age scales. SITAR (SuperImposition by Translation And Rotation) is a shape invariant growth curve model (Lindstrom, Stat Med 1995) that allows for variability not only in size, but also in terms of scale and location on the age axis. The model fits a single mean growth curve, and simultaneously estimates three random effects per subject reflecting their mean size, tempo (the timing of a growth landmark such as mean age at peak velocity) and velocity of growth (relative to the average). The random effects shift and scale each individual growth curve such that they can be superimposed on the mean curve, allowing it to be efficiently estimated as a cubic regression B-spline. The model, which is fitted in R using nlme and the author's sitar library, explains >99% of the variance of height during puberty. Examples will include height, hormones and bone maturation in puberty, and length and weight in infancy. SITAR is also useful in life course epidemiology, linking early growth to later outcome.

email: *tim.cole@ucl.ac.uk*

21. RECENT ADVANCES IN LIFETIME DATA ANALYSIS

ESTIMATION AND INFERENCE FOR THE INCREMENTAL COST-EFFECTIVENESS RATIO FOR CENSORED SURVIVAL DATA

Donna L. Spiegelman*, Harvard School of Public Health

Polyna Khudyakov, Harvard School of Public Health

Molin Wang, Harvard School of Public Health

The incremental cost-effectiveness ratio (ICER) is used to compare competing interventions for the same goal. It is a function of the expected life span (ELS) and the unit cost of treatment over the life span for each intervention compared. We develop an estimator of ICER for use in both observational cohort studies and randomized clinical trials (RCTs), that allows adjusting for any number of covariates which are determinants of the outcome of interest. These covariates are thus modifiers of the ICER even when they are not confounders of the estimated intervention effect. We also develop an estimator of the restricted ELS, that is, the ELS between two time landmarks, using a semi-parametric Cox regression model. In addition to its utility in allowing for effect modification by outcome risk factors, the proposed approach is applicable under left truncation and when the proportional hazards assumption does not apply. The method is illustrated through an analysis of the cost-effectiveness of switching from 1st to 2nd line antiretroviral drugs among 2nd line eligible patients in Dar es Salaam, Tanzania. A user-friendly SAS macro is available for implementing the method.

email: *stdls@hsph.harvard.edu*

STATISTICAL METHODS FOR RECURRENT EVENT DATA WITH MISSING EVENT CATEGORY

Jianwen Cai*, University of North Carolina, Chapel Hill

Feng-Chang Lin, University of North Carolina, Chapel Hill

Jason P. Fine, University of North Carolina, Chapel Hill

Huichuan J. Lai, University of Wisconsin, Madison

Recurrent event data frequently arise in longitudinal studies when study subjects possibly experience more than one event during the observation period. Often, such recurrent events can be categorized. However, part of the categorization may be missing due to technical difficulties or recording ignorance. If the event types are missing completely at random, then a complete case analysis may provide consistent estimates of regression parameters in certain regression models, but estimates of the baseline event rates are generally biased. Previous work on nonparametric estimation of these rates has utilized parametric missingness models. In this talk, we develop fully nonparametric methods in which the missingness mechanism is completely unspecified. Consistency and asymptotic normality of the nonparametric estimators of the mean event functions accommodate nonparametric estimators of the missingness mechanism which converge more slowly than the parametric rate. Plug-in variance estimators are provided and perform well in simulation studies, where complete case estimators may exhibit large biases and parametric estimators are generally less efficient due to model misspecification. The proposed methods are applied to the cystic fibrosis registry data.

email: *cai@bios.unc.edu*

EFFICIENT DESIGN AND ANALYSIS OF PREVALENT COHORT STUDIES

Yu Shen*, University of Texas MD Anderson Cancer Center

Hao Liu, Baylor College of Medicine

Jing Ning, University of Texas MD Anderson Cancer Center

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

The cross-sectional prevalent sampling design gives rise to length-biased data that require specialized analysis strategy but can improve study efficiency, compared to the incident sampling design. The power and sample size calculation methods are however lacking for studies with prevalent cohort design, and using the formula developed for traditional survival data may overestimate sample size. We derive the sample size formulas that are appropriate for the design of cross-sectional prevalent cohort studies, and perform numerical and simulation studies to compare the sample size requirements for achieving the same power between prevalent cohort and incident cohort designs. Using rigorous designs and proper analysis tools, the prospective prevalent cohort design can be more efficient than the incident cohort design with the same total sample sizes and study durations. This is a joint work with Hao Liu, Jing Ning and Jing Qin.

email: *yshen@mdanderson.org*

VARIABLE SELECTION FOR PENALIZED THRESHOLD REGRESSION

Xin He*, University of Maryland, College Park

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

As an alternative approach to the Cox proportional hazards model, threshold regression (TR) is a relatively new methodology that does not require a proportional hazard assumption to analyzing time-to-event data. In this talk, penalized likelihood approaches are proposed to handle the variable selection problem in the context of threshold regression analysis. The proposed methods simultaneously select significant variables and estimate unknown regression coefficients. An algorithm is presented for this process. Simulation studies are conducted for assessing the performance of the proposed approach, and the methodology is applied to a motivating study of osteoporotic fractures.

email: *xinhe@umd.edu*

22. ANALYSIS OF LONGITUDINALLY OBSERVED FUNCTIONAL DATA

A FUNCTIONAL DATA MODEL FOR ANALYZING LONGITUDINAL CHANGE OF DAILY PHYSICAL ACTIVITY

Oliver Chen, Johns Hopkins University

Luo Xiao*, North Carolina State University

Martin Lindquist, Johns Hopkins University

Jennifer Schrack, Johns Hopkins University

Luigi Ferrucci, National Institute on Aging, National Institutes of Health

Ciprian Crainiceanu, Johns Hopkins University

Objective measurement of physical activity using wearable devices such as accelerometers may provide tantalizing new insights into the association between activity and health outcomes. Accelerometers can record quasi-continuous activity information for many days and for hundreds of individuals. For example, in the Baltimore Longitudinal Study on Aging daily physical activity was recorded for about 300 adults for several days during each visit and each subject has 2 to 4 visits. An interesting problem is to quantify how daily physical activity patterns change with age, sex, body mass index and other covariates. We propose a longitudinal functional data model where the parameters of interest are bivariate functions of time and age. To deal with the complex correlation structure in the data, we use a GEE-type approach for model estimation and we propose a two-step procedure for efficient estimation. Results reveal several interesting, previously unknown daily activity patterns associated with human aging.

email: *lxiao5@ncsu.edu*

MODERN ANALYSIS OF LONGITUDINAL FUNCTIONAL DATA

So Young Park, North Carolina State University

Ana-Maria Staicu*, North Carolina State University

We propose a new modeling framework to analyze functional data that are correlated because of a longitudinal-based design: each subject is observed at repeated time visits and for each visit we record a functional variable. Our approach allows to accurately describe the process dynamics over visits, but also provides prediction of a full trajectory at a future visit. The methodology is characterized by high predictive accuracy, and yields interpretable models, while retaining computational efficiency. The proposed methodology is investigated numerically in finite samples. The method is inspired by and applied to a longitudinal diffusion tensor imaging study of multiple sclerosis, where the focus is to study the natural evolution/dynamics of the disease over time.

email: *ana-maria_staicu@ncsu.edu*

INFERRING BRAIN SIGNAL SYNCHRONICITY FROM A SAMPLE OF EEG READINGS

Donatello Telesca*, University of California, Los Angeles

Qian Li, University of California, Los Angeles

Damla Senturk, University of California, Los Angeles

Catherine Sugar, University of California, Los Angeles

Inferring patterns of synchronous brain activity from a heterogeneous sample of electroencephalograms (EEG) is scientifically and methodologically challenging. While it is statistically appealing to rely on readings from more than one individual, in order to highlight patterns of recurrent brain activation, pooling information across subjects presents with non trivial methodological problems. We discuss some of the scientific issues associated with the understanding of coordinated neuronal activity and propose a methodological framework for statistical inference from a sample of EEG readings. Our work builds on classical cotributions in time-series, cluster and functional data analysis, in an effort to reframe a challenging inferential problem in the context of familiar analytical techniques. Some attention will be paid to computational issues, with a proposal based on the hybrid combination of machine learnig and Bayesian techniques.

email: *donatello.telesca@gmail.com*

23. ADAPTIVE DESIGNS AND ADAPTIVE RANDOMIZATION

OPTIMAL AND LEAD-IN ADAPTIVE ALLOCATION FOR BINARY OUTCOMES: A COMPARISON OF BAYESIAN METHODOLOGIES

Roy T. Sabo*, Virginia Commonwealth University

Ghalib Bello, Arbor Research Collaborative for Health

Using outcome-adaptive allocation strategies for treatment assignment in clinical trials can often lead to greater successfully treated patients than using equal allocation. Such allocation procedures that incorporate Baysian estimators in many cases also maintain desired type I and II error rates. Here we compare teh use of several posterior and predictive estimators and probabilities in response-adaptive randomization designs for two- and three-group clinical trials with binary outcomes. Adaptation methods based upon posterior estimates are discussed, as are two predictive probability algorithms: one using the traditional definition, the other using a skeptical distribution. Optimal and natural lead-in designs are covered. Simulation studies show: efficacy comparisons lead to more adaptation than center comparisons, though with some power loss; skeptically predictive efficacy comparisons and natural lead-in approaches lead to less adaptation but offer reduced allocation variability. Though nuanced, these results help clarify the power-adaptation trade-off in adaptive randomization.

e-mail: *roy.sabo@vcuhealth.org*

MORE EFFICIENT TREATMENT COMPARISON IN CROSS-OVER DESIGN BY ALLOCATING SUBJECT BASED ON RANKED AUXILIARY VARIABLES

Yisong Huang*, Georgia Southern University

Hani Samawi, Georgia Southern University

In public health studies, there are many diseases hard to cure or health-risk factors hard to clearly eliminate from environment. But there are ways to moderating its effect to health. In such studies, the effect of treatments is primary interests of those experiments. The sequence in which the subjects receive treatments is not of interest. Experiments are designed in such a way that each subject, is given a number of treatments with the object of studying differences between these treatments. This experimental design, named cross-over design is widely used in clinical studies, behavioral interventions, environment experiments, epidemiology researches and animal studies. Many studies use Latin Square with cross-over design. It is used to eliminate two nuisance sources of variability. To allow inference from sampling results back to the population, simple random sample is widely used as a strategy of selection representative sample in cross-over studies. There are several desirable properties about simple random sample: easy to understand, easy to use, no selection bias, and it produces an unbiased estimator for the population mean. However, occasionally simple random sample cannot produce representative samples. Under some experimental setting, when nuisance factor is known and controllable, making an informal measurement on a unit is far cheaper than making a formal measurement, simple random sample may also become expensive. One solution to address these deficiencies in experiments is to construct an informative sampling design using available related information. Ranked set sampling is introduced to improve the efficiency of treatment comparison in cross-over design.

e-mail: *yh00049@georgiasouthern.edu*

A BAYESIAN SEQUENTIAL DESIGN WITH BINARY OUTCOME

Han Zhu*, Louisiana State University

Qingzhao Yu, Louisiana State University

Donald Mercante, Louisiana State University

We propose a Bayesian sequential design for binary outcome using an alpha spending function to control the overall type I error rate. Algorithms are presented for calculating critical values and power for the proposed designs. We also propose a new stopping rule for futility. Sensitivity analysis is implemented for assessing the effects of varying the parameters of the prior distribution and maximum total sample size on critical values. Alpha spending functions are compared using power and actual sample size through simulations. Further simulations show that, when total sample size is xed, the proposed design has greater power than the traditional Bayesian

sequential design, which sets equal stopping bounds at all interim analyses. We also nd that the proposed design with the new stopping for futility rule results in larger power and can stop earlier with a smaller actual sample size, compared with the traditional stopping rule for futility when all other conditions are hold constant. Finally, we apply the proposed method to a real data set and compare the results with traditional designs.

e-mail: *hzhu1@lsuhsc.edu*

AN EFFICIENT METHOD TO SIMULATE BAYESIAN ADAPTIVE CLINICAL TRIALS

Zhenning Yu*, Medical University of South Carolina

Viswanathan Ramakrishnan, Medical University of South Carolina

Caitlyn Ellerbe, Medical University of South Carolina

Adaptive seamless designs (ASD) may reduce development costs and shorten the drug development timeline. In ASD, accumulated information from interim data are used to update the trial design. As such, this is a natural complement to Bayesian methodology in which the prior clinical belief is sequentially updated using the observed probability of success. Simulation is often required for Bayesian ASDs due to the complexity of the design and to optimize key design characteristics. Unfortunately two limiting factor in simulations is the computational burden and time to obtain results. We present a Python module with C-extension for the design of a Bayesian ASD for Multi-Arm Multi-Stage (MAMS) Designs. The application allows users to specify a scenario of interest and returns a formatted report of the results. To address the computational time and burden, we have optimized a method for calculating the posterior predictive probability of success in a MAMS design. The resulting module significantly increases the simulation speed by implementing a series of sub-modules that have wider statistical applications. Finally, a Graphic User Interface of the simulation approach is introduced.

e-mail: *yuz@musc.edu*

MULTI-STAGE DOSE-SCHEDULE FINDING DESIGNS FOR PRE-CLINICAL STUDIES IN STROKE

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

Jing Ning, University of Texas MD Anderson Cancer Center

Xuelin Huang, University of Texas MD Anderson Cancer Center

There has been much research on adaptive dose-finding designs for clinical trials, but little has been done to improve the efficiency of identifying optimal doses in pre-clinical studies, especially when both doses and schedules need to be determined. Motivated from an animal study for stroke, we propose a Bayesian multi-stage dose-schedule finding design to simultaneously identify the optimal doses and suitable time window for the investigated drug. A piece-wise logistic model is used to accommodate the possible non-monotonic pattern for the dose-schedule-efficacy relationship and an adaptively shrinkage algorithm is developed to assign more cohorts to the potential optimal doses. The performance of the proposed design is evaluated through extensive simulation studies and compared to the standard factorial design.

e-mail: *ccai.stat@gmail.com*

THE MOST POWERFUL TEST AND THE ORDER OF ERROR PROBABILITIES FOR RESPONSE ADAPTIVE DESIGNS

Yanqing Yi*, Memorial University of Newfoundland

Xuan Li, University of Minnesota, Duluth

Response adaptive designs have the ethical advantages over the traditional methods, but the numbers of patients allocated to treatments are random due to the adaptation process of treatment allocation. This paper discusses the asymptotic optimality of statistical inference for response adaptive designs. The upper bound of statistical power of asymptotically level tests is derived and the Wald statistic is shown to be asymptotically optimal in terms of achieving the upper bound. The rates of error probability of the confidence interval and hypothesis testing are proven to depend on the convergence rate of the allocation proportions. When the convergence rate of allocation proportions is unknown, the orders of coverage error probability of confidence interval and the type I error rate are established. Specially, if the response density functions are normal density functions, these orders can be improved.

e-mail: *Yanqing.Yi@med.mun.ca*

24. CLINICAL TRIALS

STOCHASTIC MODELING OF PATIENTS RECRUITMENT IN CLINICAL TRIALS

Nicolas J. Savy*, Mathematics Institute of Toulouse

In the framework of a clinical trial, an important question to deal with is how long it takes to recruit a given number of patients. Until now, most techniques were based on deterministic models and various ad hoc techniques. Using a Poisson process to describe the recruitment process is now an accepted approach. However, in real trials, different centres have different recruitment rates. To mimic this variation, Anisimov and Fedorov (2007) introduced the so-called Poisson-gamma model in which the patients are assumed to arrive at different centres according to Poisson processes with Gamma-distributed rates. This model has been evaluated on many

real datasets and yields to relevant predictive properties. This talk aims to introduce the main ideas of Poisson-gamma recruitment modelling and its usefulness on clinical trial follow up. It is a natural question to wonder if breaks in recruitment process generate difficulties on estimation of the expected duration. This question is discussed during this talk and some highlights are given by means of a simulation study. The talk ends by a presentation of results on the use of Poisson-gamma model to test feasibility of a clinical trial recruitment given information from previous clinical trials.

email: *Nicolas.Savy@math.univ-toulouse.fr*

A MULTI-STATE MODEL FOR DESIGNING CLINICAL TRIALS FOR TESTING OVERALL SURVIVAL ALLOWING FOR CROSSOVER AFTER PROGRESSION

Fang Xia*, University of Texas MD Anderson Cancer Center

Stephen L. George, Duke University

Xiaofei Wang, Duke University

In designing a clinical trial for comparing two or more treatments with respect to overall survival (OS), a proportional hazards assumption is commonly made. However, in many cancer clinical trials, patients pass through various disease states prior to death and because of this may receive treatments other than originally assigned. For example, patients may crossover from the control treatment to the experimental treatment at progression. Even without crossover, the survival pattern after progression may be very different than the pattern prior to progression. The proportional hazards assumption will not hold in these situations and the design power calculated on this assumption will not be correct. Here we describe a simple but intuitive multi-state model allowing for progression, death before progression, post-progression survival and crossover after progression and apply this model to the design of clinical trials for comparing the OS of two treatments. For given values of the parameters of the multi-state model, we simulate the required number of deaths to achieve a specified power and the distribution of time required to achieve the requisite number of deaths. The results may be quite different from those derived using the usual PH assumption.

email: *fang.katrina.xia@gmail.com*

CONTROL OF FALSE POSITIVES IN RANDOMIZED PHASE III CLINICAL TRIALS

Changyu Shen*, Indiana University

Ziyue Liu, Indiana University

Huiping Xu, Indiana University

Hai Liu, Gilead Sciences, Inc.

Cynthia Yue, Indiana University

WITHDRAWN

Randomized Phase III clinical trials serve as the gold-standard for the evaluation of the efficacy of a medical intervention. Although research and development in earlier stages together with rigorous statistical examination assures small probability of false positive for a given trial, it is unclear how many false positives were generated from the large number of trials from the biopharmaceutical industry in the United States. The Proportion of comparisons in Phase III trials where the medical intervention has Null or Negative efficacy, or PNN, is at the central position for the estimation and control of the number of false positives. We seek to estimate PNN using a new Bayesian deconvolution method. Using data from clinicaltrials.gov (CT.gov) and other data sources, we identified 1393 trials that meet our study entry criteria, which are dominated by trials on drugs for treatment purpose. Among the 1221 trials with result available on selected comparison, 789 (64.6%) show statistically significant superiority of the intervention, with 561 (45.9%) having a two-sidedp-value less than 0.001. The PNN is estimated to be no more than 7-9%, leading to an expectation of no more than 6-8 trials with at least one false positive comparison over a 5-year period.

email: *chashen@iu.edu*

ONE-SIDED GLOBAL TESTS FOR MULTIVARIATE OUTCOMES IN RANDOMIZED TRIALS

Donald Joseph Hebert*, University of Rochester Medical Center

Global tests are highly useful in randomized trials in diseases with multiple outcome measures of equal importance. Such methods play a pivotal role in assessing an overall treatment effect and are particularly powerful in the case where the treatment effects on individual outcomes are consistent, i.e., in the same direction. Much attention has been given to this problem when the outcomes are assumed to follow a multivariate normal distribution. O'Brien's tests (1984) based on ordinary and generalized least squares are very powerful when the individual effects are similar in magnitude. When the effects are dissimilar, these procedures lack power compared to other tests such as the approximate likelihood ratio test (Tang et al., 1989, Glimm et al., 2002, Tamhane and Logan, 2002). This talk will present a novel class of global tests based on procedures for combining p-values derived from orthogonalized test statistics. Simulation studies demonstrate that these tests can provide high power that is stable across different alternatives in the positive orthant. Outcome-specific inference and extensions to outcomes of mixed type will also be discussed.

email: *Donald_Hebert@urmc.rochester.edu*

INEQUALITY IN TREATMENT BENEFITS: CAN WE DETERMINE IF A NEW TREATMENT BENEFITS THE MANY OR THE FEW?

Emily J. Huang*, Johns Hopkins University

Ethan X. Fang, Princeton University

Michael A. Rosenblum, Johns Hopkins University

The primary analysis in many randomized controlled trials focuses on the average treatment effect and does not address 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. Our goal is to learn about the fraction who benefit from a treatment, based on randomized trial data. We consider the case where the outcome is ordinal, with binary outcomes as a special case. In general, the fraction who benefit is a non-identifiable parameter, and the best that can be obtained are sharp lower and upper bounds on it. Our main contributions include (i) showing that the naive (plug-in) estimator of the bounds can be inconsistent if support restrictions are made on the joint distribution of the potential outcomes (such as the no harm assumption); (ii) developing the first consistent estimator for this case; (iii) applying this estimator to a randomized trial dataset of a medical treatment to determine whether the estimates can be informative. Our estimator can be computed using linear programming, allowing fast implementation.

email: *emhuang1@gmail.com*

FACTORIAL CLINICAL TRIALS FOR HYBRID RESEARCH STUDIES: DESIGN AND ANALYSIS OF OPTIMIZING TREATMENT FOR COMPLICATED GRIEF

Christine M. Mauro*, Columbia University

Xin Qiu, Columbia University

Donglin Zeng, University of North Carolina, Chapel Hill

Naihua Duan, Columbia University

Yuanjia Wang, Columbia University

Complicated grief is a psychiatric disorder that affects many patients for long durations, with substantial impact on quality of life, functioning, and productivity. A psychotherapy, complicated grief treatment (CGT) has been shown to be efficacious in two previous randomized clinical trials. However there is an important need to assess efficacy for medication (citalopram), either as a stand-alone treatment, or in conjunction with CGT. Of primary interest is comparing: Aim 1. Citalopram vs. Placebo (without CGT) and Aim 2. Citalopram + CGT vs. Placebo + CGT. Of secondary interest is to assess benefits for CGT under naturalistic conditions, especially taking into consideration medication status, that is, Aim 3. Citalopram + CGT vs. Citalopram + no CGT and Aim 4. CGT vs. no CGT (no Citalopram). Finally, there is also interest in understanding the interaction between "Citalopram vs. Placebo" vs. "CGT vs. no CGT." In order to efficiently address all of these aims, a 2x2 factorial randomized trial was designed, implemented and is now completed. In this talk, we present analyses methods for examining the primary aims and new subgroup analysis methods based on examining qualitative interactions.

email: *cmm2212@cumc.columbia.edu*

ESTIMATING INDIVIDUALIZED TREATMENT RULES FOR ORDINAL TREATMENTS

Jingxiang Chen*, University of North Carolina, Chapel Hill

Yufeng Liu, University of North Carolina, Chapel Hill

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

Haoda Fu, Eli Lilly and Company

Xuanyao He, Eli Lilly and Company

Precision medicine is an emerging scientific approach for disease treatment and prevention by taking into account individual variability of a patient. It is an important direction for clinical trial research and many statistical methods have been proposed recently. One of the primary goals in precision medicine is to obtain an optimal individual treatment rule (ITR) which can help make decisions on treatment selection according to each patient's specific characteristics. Recently, outcome weighted learning (OWL) has been proposed to estimate such an optimal ITR in the binary treatment setting by maximizing the expected clinical outcome. However, for the ordinal treatment settings such as dose finding, it is unclear how to use OWL. Furthermore, OWL requires transformation of the clinical outcome when the outcome has negative values. In this paper, we propose a new technique for estimating ITR with ordinal treatments. In particular, we propose a data duplication technique with a piecewise convex loss function. We establish Fisher consistency for the resulting estimated ITR under certain conditions. We also obtain the convergence and risk bound properties. Simulated examples and an application on an irritable bowel example demonstrate the highly competitive performance of the proposed method compared to several existing ones.

email: *jgxchen@email.unc.edu*

25. CLUSTERED DATA METHODS

LEARNING PARAMETER HETEROGENEITY IN DATA INTEGRATION

Lu Tang*, University of Michigan

Peter X.K. Song, University of Michigan

As data sets of related studies become more easily accessible, combining data sets of similar studies is undertaken in practice to achieve a larger sample size and higher power. A major challenge arising from data integration pertains to data heterogeneity in terms of population, study coordination, or experimental protocols. Ignoring such heterogeneity in data analysis may result in biased estimation and misleading inference. Traditional techniques of remedy to data heterogeneity include the use of interactions and random effects, which are inferior to achieving desirable statistical power or providing an intuitive interpretation, especially when a large number of smaller data sets are combined. In this paper, we propose a regularized fusion method that allows us identify and merge inter-study homogeneous parameter clusters in regression analysis, without the use of hypothesis testing approach. Using fused lasso, we establish a computationally efficient procedure to deal with large-scale integrated data. Incorporating the estimated parameter ordering in the fused lasso facilitates computing speed with no loss of statistical power. We conduct extensive simulation studies and provide an application example to demonstrate the performance of the new method with a comparison to the conventional methods.

email: *lutang@umich.edu*

CLUSTERS WITH RANDOM SIZE: WEIGHTED ESTIMATION FOR COMPOUND SYMMETRY AND AR(1) MODELS

Lisa Hermans*, Universiteit Hasselt, Belgium

Vahid Nassiri, Katholieke Universiteit Leuven, Belgium

Geert Molenberghs, Universiteit Hasselt and Katholieke Universiteit Leuven, Belgium

Michael G. Kenward, London School of Hygiene and Tropical Medicine

Wim Van der Elst, Universiteit Hasselt, Belgium

Marc Aerts, Universiteit Hasselt, Belgium

Geert Verbeke, Katholieke Universiteit Leuven and Universiteit Hasselt, Belgium

There are many contemporary statistical designs that do not use a random sample of a fixed, a priori determined size. Examples include settings with clusters of random size, in the literature often referred to as “informative cluster sizes”. While the latter means that the cluster size is associated with the data values in the cluster, attention here is confined to issues that arise even when the data and the cluster size are unrelated. With unequal cluster sizes useful features such as the existence of complete sufficient statistics are no more available (Hermans et. al. 2015), whereas subsamples with clusters of equal size do obtain closed-form solutions and complete sufficient statistics. This suggested the use of sample splitting, a pseudo-likelihood based approach that allows the data to be analyzed ‘by parts’ (Molenberghs et al, Stat Probab Lett, 2011). The results are then combined using suitable weights. Two settings are studied in detail: the compound symmetry and AR(1) covariance structure. Several weights are compared and findings are illustrated using data from a developmental toxicity study, where clusters are formed of fetuses within litters, and data repeatedly measuring Positive and Negative Syndrome Scale for in psychiatric patients.

email: *lisa.hermans@uhasselt.be*

GOODNESS OF FIT TEST FOR MULTINOMIAL REGRESSION MODEL IN NUN STUDY

Zhiheng Xie*, University of Kentucky

Richard Kryscio, University of Kentucky

Discrete-time Markov chains have been used to analyze the transition of subjects from intact cognition to dementia with mild cognitive impairment and global impairment as intervening transient states, and death as competing risk. A multinomial logistic regression model is used to estimate the probability distribution in each row of the one step transition matrix that correspond to the transient states. We investigate some goodness of fit tests for a multinomial distribution with covariates to assess the fit of this model to the data. We propose a modified chi-square test statistic and a score test statistic for the multinomial assumption in each row of the transition probability matrix. We apply the test to the data from the Nun Study, a cohort of 461 participants. We incorporate presence or absence of the APOE-4 allele, education, and age as covariates in the application.

email: *zhiheng.xie@uky.edu*

SEQUENTIAL IMPUTATION USING MARGINAL MODELS

Recai M. Yucel*, State University of New York, Albany

Zeynep I. Kalaylioglu, Middle East Technical University

Skip patterns, bounds and diverse measurement scales often exacerbate the problem of item nonresponse in the analysis of survey data. Variable-by-variable imputation techniques have been quite successfully applied to conduct inference by multiple imputation to overcome such problems. Most of the methods so far utilized conditional models to sample from the approximate posterior predictive distributions of missing data. MI inferences under these models perform quite well when the goal of the inferences target conditional quantities. We argue that when the goal of the inference is population- averaged quantities such as fixed-effects, marginal models fitted using generalized estimating equations can be attractive alternatives. We pursue the sampling from the corresponding predictive distributions of missing data using bootstrap methods. These methods are readily available in many statistical software and do not require any advanced computational technique allowing many practitioners to easily adopt them. In our talk, we will present

a comprehensive simulation study demonstrating the proposed method's repetitive sampling properties along with a data example.

email: *ryucel@albany.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*

JOINT CLUSTERING AND INFERENCE IN FUNCTIONAL DATA PROTEIN SPECTROSCOPIC PROFILES: APPLICATIONS IN THE EYE LENS PROTEIN CRYSTALLIN

Miranda L. Lynch*, University of Connecticut Health Center

In many applications of clustering of functional data, a primary goal is to locate subgrouping of profiles that assign individual curves to clustered subsets. Less work has been carried out in situations where some elements of grouping structure are partially known, and in which the nature of the curve separation is itself of interest. In this work, we present combining hierarchical clustering of functional profiles with inference on the nature of curve separation. The area of application is in investigating the eye lens protein crystallin and its role in cataract formation and other disease states. The functional data methods presented here are developed for protein spectroscopic measurements using fluorescence and circular dichroism spectroscopy applied to a series of crystallin variant forms potentially relevant to cataract formation. The methods are useful for both separating the spectroscopic profiles, in addition to characterizing important profile features that represent relevant physiological states. We investigate the methods using data on a series of mutational forms of crystallin as well as simulated data to examine the ability of the methods to capture curve separation and identification of different functional forms.

email: *mlynch@uchc.edu*

MIXTURE MODELING FOR LONGITUDINAL DATA

Xiwei Tang*, University of Illinois, Urbana-Champaign

Annie Qu, University of Illinois, Urbana-Champaign

In this paper, we propose an unbiased estimating equation approach for a two-component mixture model with correlated response data. We adapt the mixture-of-experts model and a generalized linear model for component distribution and mixing proportion, respectively. The new approach only requires marginal distributions of both component densities and latent variables. We utilize serial correlations from subjects' subgroup memberships, which improves estimation efficiency and classification accuracy, and show that estimation consistency does not depend on the choice of the working correlation matrix. The proposed estimating equation is solved by an Expectation-Estimating-Equation (EEE) algorithm. In the E-step of the EEE algorithm, we propose a joint imputation based on the conditional linear property for the multivariate Bernoulli distribution. In addition, we establish asymptotic properties for the proposed estimators and the convergence property using the EEE algorithm. Our method is compared to an existing competitive mixture model approach in both simulation studies and an election data application.

email: *xtang14@illinois.edu*

26. HIGH DIMENSIONAL MODELING AND INFERENCE

PROVABLE SMOOTHING APPROACH IN HIGH DIMENSIONAL GENERALIZED REGRESSION MODEL

Fang Han, Johns Hopkins University

Honglang Wang*, Indiana University-Purdue University, Indianapolis

The generalized regression model is an important semiparametric generalization to the linear regression model. It assumes there exist unknown monotone increasing link functions connecting the response Y to a single index of d explanatory variables. The generalized regression model covers a lot of well-exploited statistical models. It is appealing in many applications where regression models are regularly employed. In low dimensions, rank-based M-estimators are recommended, giving root-n consistent estimators of the coefficients. However, their applications to high dimensional data are questionable. This is mainly due to the discontinuity of loss function. In detail, (i) computationally, because of the non-smoothness of the loss function, the optimization problem is intractable; (ii)

theoretically, the discontinuity of the loss function renders difficulty for analysis in high dimensions. In contrast, this paper suggests a simple, yet powerful, smoothing approach for rank-based estimators. A family of smoothing functions is provided, and the amount of smoothing necessary for efficient inference is carefully calculated. We show the resulting estimators are scaling near-optimal, i.e., they are consistent estimators of the coefficients as long as (n,d,s) are within a near-optimal range (Here s represents the sparsity degree). These are the first such results in the literature. The proposed approaches' power is further verified empirically.

email: *hlwang@iupui.edu*

CONFIDENCE INTERVALS FOR HIGH-DIMENSIONAL LINEAR REGRESSION: MINIMAX RATES AND ADAPTIVITY

Tony Cai, University of Pennsylvania

Zijian Guo*, University of Pennsylvania

Confidence sets play a fundamental role in statistical inference. In this paper, we consider confidence intervals for high dimensional linear regression with random design. We first establish the convergence rates of the minimax expected length for confidence intervals in the oracle setting where the sparsity parameter is given. The focus is then on the problem of adaptation to sparsity for the construction of confidence intervals. Ideally, an adaptive confidence interval should have its length automatically adjusted to the sparsity of the unknown regression vector, while maintaining a prespecified coverage probability. It is shown that such a goal is in general not attainable, except when the sparsity parameter is restricted to a small region over which the confidence intervals have the optimal length of the usual parametric rate. It is further demonstrated that the lack of adaptivity is not due to the conservativeness of the minimax framework, but is fundamentally caused by the difficulty of learning the bias accurately.

email: *zijguo@wharton.upenn.edu*

IMPROVED ESTIMATION FOR HIGH DIMENSIONAL GENERALIZED LINEAR MODELS

Abhishek Kaul, National Institute of Environmental Health Sciences, National Institutes of Health

Akshita Chawla*, Merck Research Laboratories

Tapabrata Maiti, Michigan State University

In this paper we investigate estimation via MLE post model selection in the context of high dimension generalized linear models. We show that by separating model selection and estimation, it is possible to achieve improved convergence rates of the l2 estimation error in comparison to simultaneous estimation and variable selection methods such as l1 penalized likelihood, i.e., faster than sqrt(s log p/n). The estimation is done via MLE on the model selected by a generic procedure. Here s, p are the number of non zero parameters and the model dimension respectively, n is the sample size. We show that under very general model selection criteria, the proposed method is at least as efficient as l1 penalized methods. We also provide convergence rates of the estimation error in the sup norm which attain a rate of sqrt(log s/n) under perfect model selection. To the best of our knowledge this is the fastest available convergence rate in the literature that holds for the sup norm of the entire p-dimensional vector simultaneously w.p. ? 1 . All results are supported empirically by a simulation study. Lastly, we implement the proposed methodology on the breast cancer data of Veer et. al. (2002).

email: *akshita.chawla@merck.com*

HIGH-DIMENSIONAL INFERENCE FOR COX MODEL

Ethan X. Fang*, Princeton University

Yang Ning, Princeton University

Han Liu, Princeton University

This paper considers the problem of hypothesis testing and confidence intervals in high dimensional proportional hazards models. Motivated by a geometric projection principle, we propose a unified likelihood ratio inferential framework, including score, Wald and partial likelihood ratio statistics for hypothesis testing. Without assuming model selection consistency, we derive the asymptotic distributions of these test statistics, establish their semiparametric optimality, and conduct power analysis under Pitman alternatives. We also develop new procedures to construct pointwise confidence intervals for the baseline hazard function and baseline survival function. Simulation studies show that all proposed tests perform well in controlling Type I errors. Moreover, the partial likelihood ratio test is empirically more powerful than the other tests. The proposed methods are illustrated by an example of gene expression dataset.

email: *xingyuan@princeton.edu*

ON LONGITUDINAL GAUSSIAN GRAPHICAL MODELS: ESTIMATION AND ASYMPTOTIC INFERENCE

Quanquan Gu*, University of Virginia

Yuan Cao, Princeton University

Yang Ning, Princeton University

Han Liu, Princeton University

We propose a new family of semiparametric graphical models for analyzing multivariate longitudinal data. In particular, we model the joint distribution of d variables across m subjects by assuming that the distribution of each subject is Gaussian with a subject-specific mean parameter and a common precision matrix

which encodes the graph. For graph estimation, we propose a novel parameter estimation method based on the QR transformation of the data. We show that such a procedure is invariant to the subject-specific component and attains the optimal parametric rates of convergence for precision matrix estimation under different norms. For uncertainty assessment, we develop a unified inferential framework including score, Wald and likelihood ratio statistics to test the presence of an specific edge. We also propose a global test to evaluate whether a given graph is the supergraph of the truth. The theoretical properties of the proposed inferential methods are established and illustrated by thorough numerical experiments.

email: *qg5w@virginia.edu*

A GENERAL THEORY OF HYPOTHESIS TESTS AND CONFIDENCE REGIONS FOR SPARSE HIGH DIMENSIONAL MODELS

Yang Ning*, Princeton University

Han Liu, Princeton University

We consider the problem of uncertainty assessment for low dimensional components in high dimensional models. Specifically, we propose a novel decorrelated score function to handle the impact of high dimensional nuisance parameters. We consider both hypothesis tests and confidence regions for generic penalized M-estimators. Compared to most existing inferential methods which are tailored for individual models, our main contribution is to develop a general framework for high dimensional inference and is applicable to a wide range of applications. From the testing perspective, we develop general theorems to characterize the limiting distributions of the decorrelated score test statistic under both null hypothesis and local alternatives. These results provide asymptotic guarantees on the type I errors and local powers of the proposed test. Furthermore, we show that the decorrelated score function can be used to construct point estimators that are semiparametrically efficient and the optimal confidence regions.

email: *y4ning@uwaterloo.ca*

27. PREDICTION AND PROGNOSTIC MODELING

PREDICTING ALZHEIMER'S DISEASE WITH BIVARIATE MIXTURE MODELING

Frank Appiah*, University of Kentucky

Erin Abner, University of Kentucky

David Fardo, University of Kentucky

Glen Mays, University of Kentucky

Richard Charnigo, University of Kentucky

Alzheimer's Disease (AD) is a progressive neurodegenerative disorder that usually results in severely impaired cognition. Approximately 5 million Americans, predominantly older adults, currently have dementia due to AD. Although there are abundant research activities to aid in early detection of Alzheimer's Disease, very little literature has employed mixture modeling to address the problem. We predict the future disease status of currently cognitively normal people using bivariate mixture modeling of ratios derived from established biomarkers (ptau181p, abeta142 and tau) from the Alzheimer's Disease Neuroimaging Initiative (ADNI) study. The number of risk strata was selected using information theoretic criteria. Posterior probabilities of belonging to the various risk strata were also obtained for each participant. The risk of developing mild cognitive impairment (MCI) /AD was assessed with Cox modeling based on these posterior probabilities. Information criteria favored modeling with three risk strata. Estimated posterior probabilities were significant in the Cox model (p-value<0.01) even in a version which adjusted for potential confounders (p-value<0.05). The c-statistic in the latter version was 73%. The ratios of biomarkers tau/abeta142 and ptau181p/abeta142 may be useful in predicting the potential of persons to develop MCI or AD while they are still cognitively normal.

email: *frank.appiah@uky.edu*

TIME-DEPENDENT PREDICTIVE ACCURACY CURVE UNDER MARKER-DEPENDENT SAMPLING

Zhaoyin Zhu*, New York University

Xiaofei Wang, Duke University

Paramita Saha Chaudhuri, McGill University

Evaluating the accuracy of a candidate biomarker signalling the onset of disease or disease condition is essential for medical decision making. A good marker would accurately identify the patients who are likely to die at a particular time in the future or who are in urgent need for active treatments. To assess the performance of a biomarker, receiver operating characteristic (ROC) curve and the area under the ROC curver (AUC) are commonly used. In many cases, the standard simple random sampling design (SRS) used for biomarker validation studies is often costly and inefficient. In order to improve the efficiency and reduce the cost of biomarker validation, marker-dependent sampling (MDS) has been advocated in which the patients selected to assess true survival time are dependent on the result of a biomarker assay. We introduce a nonparametric estimator for time-dependent AUC under MDS design. The consistency and the asymptotic normality of the proposed estimator is established. Simulation shows the unbiasedness of the proposed estimator and a significant efficiency gain of MDS design over SRS design.

email: *zhaoyin.zhu@nyu.edu*

ESTIMATING THE IMPACT OF BASING TREATMENT DECISIONS ON MARKERS THAT PREDICT RISK

Marshall D. Brown*, Fred Hutchinson Cancer Research Center

Holly Janes, Fred Hutchinson Cancer Research Center

Models that predict the risk of a clinical event are often used to guide treatment decisions: a model identifying high risk subjects can be used to recommend an intensive or experimental treatment to them. Many statistical measures have been developed to evaluate the predictive accuracy of such models, but they fail to directly capture the impact of using the marker on the population rate of clinical events or treatment-associated toxicities. We propose a method to estimate the impact of adopting marker based treatment decisions into clinical practice using data from a cohort of subjects treated with standard of care paired with data about the efficacy of the recommended treatment. We illustrate our methods with data used to identify women at high risk for epithelial ovarian cancer who can be recommended prophylactic fallopian tube removal at the time of hysterectomy.

email: *mdbrown@fredhutch.org*

BAYESIAN INFERENCE FOR BLACK HISPANIC BREAST CANCER SURVIVAL DATA

Hafiz Khan*, Texas Tech University

We used statistical probability models for breast cancer survival data for race and ethnicity. Data was collected from breast cancer patients diagnosed in United States during the years 1973–2009. We selected a stratified random sample of Black Hispanic female patients from the Surveillance Epidemiology and End Results (SEER) database to derive the statistical probability models. We used three common model building criteria which include Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), and Deviance Information Criteria (DIC) to measure the goodness of fit tests and it was found that Black Hispanic female patients survival data better fit the exponentiated exponential probability model. A novel Bayesian method was used to derive the posterior density function for the model parameters as well as to derive the predictive inference for future response. Markov Chain Monte Carlo (MCMC) method was used for obtaining the summary results of posterior parameters. Additionally, we reported predictive intervals for future survival times. The findings would be of great significance in treatment planning and healthcare resource allocation.

email: *hafiz.khan@ttuhsc.edu*

DESIGN AND ANALYSES OF TWO-PHASE STUDIES FOR PREDICTING BINARY OUTCOMES

Xinglei Chai*, University of Pennsylvania

Jinbo Chen, University of Pennsylvania

The two-phase study design is a cost effective option for assessing the predictiveness of emerging risk predictors. In this design, data on the outcome and inexpensive predictors is collected for all study subjects (Phase I), but costly predictors are only measured on a judiciously selected subset (Phase II). Statistical methodology research for two-phase designs has mainly been focused on the estimation of association parameters that describe the relationship between outcome and predictors, as well as the exploration of sampling strategies for selecting Phase II subjects to improve the estimation efficiency. Our work, instead, is focused on efficiently estimating the statistics for quantifying the value of risk predictors in predicting binary outcomes. The estimation of these statistics involves not only the association parameters, but also the risk predictor distribution. We developed general statistical methods for estimating various predictive accuracy statistics, and showed that our proposed estimators are consistent and asymptotically normally distributed through theoretical derivation. We also studied different sampling strategies for selecting Phase II subjects to improve the efficiency of estimating these predictive accuracy measures.

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

COMBINING MULTIPLE BIOMARKERS LINEARLY TO MAXIMIZE THE PARTIAL AREA UNDER THE ROC CURVE

Qingxiang Yan*, University of Texas MD Anderson Cancer Center

Leonidas E. Bantis, University of Texas MD Anderson Cancer Center

Ziding Feng, University of Texas MD Anderson Cancer Center

It is common to measure multiple biomarkers on an individual and make clinical decisions based on combinations of those biomarkes. This study considers combining multiple biomarkers linearly to maximize the partial area under the ROC curve (pAUC). Both parametric and non-parametric methods have been developed in previous studies but with limitations. Under the normality assumption, our proposed approach adopts an alternative analytic expression of the pAUC which is easy to implement. Computational considerations are given to better minimize the risk of multiple maxima. Simulation results show that our proposed approach yields overall the most accurate estimations among existing methods. When the normality assumption does not hold, a smooth kernel-based approach is presented. Simulation results suggest that the kernel-based approach have performance comparable with grid-search when combining three or less biomarkers; and it can be applied to a larger number of markers without difficulty while grid-search becomes impossible. When combining biomarkers by logistic regression has become common practice, we

have identified scenarios where logistic regression fails to maximize the pAUC. Therefore, the proposed approaches are recommended when only the high specificity/sensitivity region rather than the whole region is clinically relevant. Finally, the proposed approaches are illustrated on a prostate cancer dataset.

email: *gyan@mdanderson.org*

BUILDING BETTER GENE SIGNATURES WITH RANK-BASED FEATURES AND META-ANALYSIS

Prasad Patil*, Johns Hopkins University

Jeffrey T. Leek, Johns Hopkins University

Apart from a handful of success stories, translation of gene signatures from research to the clinic has been slower than desired. One primary cause of this is difficulty in the interpretation of the underlying predictive model that maps gene expression measurements to an outcome prediction. Another major issue is the lack of reproducibilty plaguing many facets of the signature-building process, including but not limited to biases in test set predictions due to data pre-processing and lack of generalizability beyond specific validation sets. In this work, we propose a complete signature-building procedure that is entirely reproducible. Reproducibility is achieved by minimizing user input and leveraging report-generating mechanisms like Rmarkdown and knitr to document the signature-building process from start to finish. We will describe the novel underlying feature-selection techniques which build interpretable prediction models using rank-based Top Scoring Pair (TSP) features. We will also show how using rank-based features allows us to apply the procedure across many datasets and achieve greater stability and reliability in the proposed gene signatures. We predict recurrence probability using large curated sets of microarray data from ovarian and breast cancer experiments to motivate the efficacy of our approach.

email: *prpatil42@gmail.com*

28. STATISTICAL METHODS FOR ADDRESSING CHALLENGES IN MICROBIOME AND METAGENOMIC ANALYSIS

STATISTICAL ANALYSIS OF DATA IN METRIC SPACE WITH APPLICATIONS IN MICROBIOME STUDIES

Hongzhe Li*, University of Pennsylvania

Next-generation sequencing technologies allow 16S ribosomal RNA gene surveys or whole metagenome shotgun sequencing in order to characterize taxonomic and functional compositions of gut microbiomes. The outputs from such studies are short sequence reads derived from a mixture of genomes of different species in a given microbial community. By placing these reads on a phylogenetic tree, we can treat the data as probability masses on branches of the tree with a well-defined Kantorovich-Rubinstein distance measure that accounts for the phylogenetic relationship among the species. We consider the regression problem for data in such a metric space in order to build a model for predicting the clinical outcomes based on microbiome composition. The procedure is based on a set of partitions of the metric data induced by a hierarchy of r-nets of the input data. It then applies local kernel regression based on such r-nets. The resulting smooth and adaptive regression can be used to predict the outcome. We compare the performance of this approach with the mixed-effect model approach using simulations and show improved prediction performances. We apply this method to predict BMI based on 16S data and to predict Crohn's disease based on shotgun metagenomic sequencing data. We show that using the phylogenetic tree information significantly improves the prediction.

email: *hongzhe@upenn.edu*

HIGH-PRECISION MICROBIAL COMMUNITY FUNCTIONAL PROFILING AND META'OMIC INTEGRATION

Curtis Huttenhower*, Harvard School of Public Health

Sequencing-based studies of the human microbiome have become a powerful tool for surveying whole-community microbial ecology, but their mechanistic interpretation remains challenging both computationally and biologically. Moreover, multiple high-throughput functional profiling technologies, including metabolomics and transcriptomics, are now mature and cost-effective enough to apply longitudinally to human cohorts. I will discuss recent bioinformatic approaches to data integration during human microbiome studies and their application to functional profiling and downstream epidemiological analysis. These include complementary hierarchical Bayesian models of taxonomic profiling data, microbial ecological interactions, and microbiome longitudinal covariation. I will highlight applications of these models in integrating metagenomic, metatranscriptomic, and metabolomic profiles of the gut during inflammatory bowel disease as part of the NIH Integrative Human Microbiome Project.

email: *chuttenh@hsph.harvard.edu*

SOME CHALLENGES IN THE ANALYSIS OF MICROBIOME DATA: OLD WINE IN A NEW BOTTLE WITH A TWIST!

Abhishek Kaul, National Institute of Environmental Health Sciences, National Institutes of Health

Siddhartha Mandal, Public Health Foundation of India, Gurgaon, India

Shyamal D. Peddada*, National Institute of Environmental Health Sciences, National Institutes of Health

Understanding differences in the microbial composition and abundance of taxa in different groups or populations (e.g. pre-term and full term babies) is of great interest. For a given specimen (e.g. fecal sample) obtained from an ecosystem (e.g. infants gut), one typically observes the abundance of operational taxonomic units (OTUs), which are microbial genomic sequences clustered by sequence similarity. These OTUs are then typically mapped to a taxonomic reference database to obtain an estimate of the abundance of each OTU in the specimen and not in the ecosystem. Using the observed specimen level OTU count data one can at most estimate the relative abundance of a taxon in the ecosystem but not the actual abundance in the ecosystem. Since the sum of the relative abundances of taxa is 1, they are constrained by a simplex. In this talk we describe various statistical parameters associated with the microbiome count data and describe our methodology to compare taxa abundance in two or more populations while adjusting for covariates. Additionally, we introduce methodology for describing networks among taxa. The proposed methodology accounts for inflated zero counts, a common feature of these data. The proposed methodology is illustrated using some recently published gut microbiome data.

email: *peddada@niehs.nih.gov*

FLEXIBLE METHODS FOR TESTING MICROBIOME BY ENVIRONMENT INTERACTIONS

Michael C. Wu*, Fred Hutchinson Cancer Research Center

Advances in high throughput sequencing have enabled studies of microbiome composition. Increasing, such studies are being conducted to understand how the microbiome influences clinical outcomes and response to environmental exposures. Recently, there has been considerable interest in understanding the role that microbiome plays in modifying the relationship between the exposures (broadly defined) and outcomes. However, there has been little work on testing the statistical interactions with microbiome composition. Therefore, we propose a new strategy for testing the interaction between the microbiome and an exposure using the semi-parametric kernel machine framework to jointly model the effect of the microbiome, environment, and their interactions. By exploiting the connection between mixed models and kernel methods, we employ a variance component score test. In contrast to existing applications kernel methods, the complexity of the data and strong main effects necessitate significant adaptation to correctly capture the null model in an unbiased manner and enable type I error control. Simulations and real data applications are used to demonstrate our methodology.

email: *mcwu@fhcrc.org*

29. RECENT ADVANCES AND CHALLENGES IN ADAPTIVE DESIGN FOR CLINICAL TRIALS

CONTINUAL REASSESSMENT METHOD WITH MULTIPLE TOXICITY CONSTRAINTS FOR LATE ONSET AND CUMULATIVE TOXICITIES

Shing M. Lee*, Columbia University

The toxicity profile of newer anticancer treatments such as targeted and immunotherapeutic agents differs from that of chemotherapy. While some of these newer agents cause dose limiting toxicities, others cause lower grade toxicities which may not occur within the first cycle of treatment. Thus, in the early development of these agents it is necessary to account for both lower grade toxicity, as well as, late onset and cumulative toxicity. However, methods for ordinal toxicity that can account for lower grade toxicities require complete follow-up data from patients before the next dose can be assigned. If the follow-up time to include late onset and cumulative toxicities is long, it will have an impact on time to completion of these studies and delay middle and late development. We propose an extension to the continual reassessment method with multiple constraints that can accommodate for incomplete follow-up data. This method allows for patients to be entered before complete follow-up is observed and can impose constraints on milder toxicities. We present the method in the context of cancer clinical trial.

email: *sml2114@columbia.edu*

SEQUENTIAL DESIGN METHOD FOR BIOEQUIVALENCE TEST WITH SERIAL SAMPLING DATA

Fangrong Yan*, Pharmaceutical University, China

Junling Liu, Pharmaceutical University, China

Xueling Huang, University of Texas MD Anderson Cancer Center

The planning of a bioequivalence test requires an assumption about the variance that is used to estimate the sample size. This assumed variance may be too small, which leads to an underestimated sample size and underpowered study. This problem is magnified in a bioequivalence test with serial sampling data. In serial sampling data only one sample is collected from each individual and the correct assumption of variance becomes even more difficult. To solve this problem, we apply sequential design methods to bioequivalence test with serial sampling data. We propose four 3-stage sequential designs in contrast to 2-stage sequential designs. The proposed 3-stage designs are expected to increase the power and reduce the sample size needed for the bioequivalence test compared to that for the 2-stage designs. Simulations are conducted to show the power and type I error rate for each method. The results show that the performances of all methods proposed herein are similar when

the variability is small, and the proposed 3-stage sequential design methods outperform the 2-stage sequential design methods when the variability is large.

email: *f.r.yan@163.com*

BAYESIAN OPTIMAL INTERVAL (BOIN) DESIGNS FOR PHASE I CLINICAL TRIALS

Ying Yuan*, University of Texas MD Anderson Cancer Center

Suyu Liu, University of Texas MD Anderson Cancer Center

In phase I trials, effectively treating patients and minimizing the chance of exposing them to subtherapeutic and overly toxic doses are clinician's top priority. Motived by this practical consideration, we propose Bayesian optimal interval (BOIN) designs to find the maximum tolerated dose (MTD) and minimize the probability of inappropriate dose assignments for patients. We show, both theoretically and numerically, that the BOIN design not only has superior finite- and large-sample properties, but also can be easily implemented in a simple way similar to the traditional 3+3 design. Compared to the well-known continual reassessment method, the BOIN design yields comparable average performance to select the MTD, but has a substantially lower risk of assigning patients to subtherapeutic and overly toxic doses.

email: *yyuan@mdanderson.org*

PHASE I-II CLINICAL TRIALS WITH DELAYED OUTCOMES

Joseph S. Koopmeiners*, University of Minnesota

The primary objective of phase I oncology trials is to identify the maximum tolerated dose (MTD), defined as the maximum dose with probability of dose limiting toxicity less than some pre-specified threshold. The results of phase I are used to identify the dose that will be evaluated for efficacy in Phase II. An alternate approach is to combine Phases I and II into a single trial that considers the trade-off between efficacy and toxicity during dose-finding. A practical limitation to implementing these designs, referred to as Phase I-II designs, is the timely availability of both outcomes. For example, toxicity is typically evaluated over a single course of treatment, while efficacy may not be evaluated for several months. As a result, new patients may be ready to enroll in the trial before the outcomes of the previous patients have been observed. In this talk, we will discuss two approaches to phase I-II clinical trials with delayed outcomes. In the first approach, we treat efficacy and toxicity as time-to-event outcomes and in the second, we incorporate a surrogate endpoint for efficacy that is available prior to measuring the primary efficacy endpoint.

email: *koopm007@umn.edu*

30. HEALTH CARE PROVIDER EVALUATION

A DIRICHLET PROCESS MIXTURE MODEL FOR SURVIVAL OUTCOME DATA: ASSESSING NATIONWIDE KIDNEY TRANSPLANT CENTERS

Lili Zhao, University of Michigan

Jing Chunzi Shi, University of Michigan

Tempie Shearon, University of Michigan

Yi Li*, University of Michigan

Mortality rates are probably the most important indicator for the performance of kidney transplant centers. Motivated by the national evaluation of mortality rates at kidney transplant centers in the United States, we seek to categorize the transplant centers based on the mortality outcome. We describe a Dirichlet process model and a Dirichlet process mixture model with a half-cauchy prior for the estimation of the risk-adjusted effects of the transplant centers, with strategies for improving the model performance, interpretability as well as classification ability. We derive statistical measures and create graphical tools to rate transplant centers and identify outlying groups of centers with exceptionally good or poor performance. The proposed method was evaluated through simulation, and then applied to assess kidney transplant centers from a national organ failure registry.

email: *yili@umich.edu*

HEALTHCARE PROVIDER COMPARISONS: IDENTIFYING AND MEETING GOALS

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

The Centers for Medicare and Medicaid Services (CMS) annually compares hospitals with respect to mortality and readmissions. Metrics are computed by comparing hospital-specific performance to that for a counterfactual hospital treating the same patients, but operating at the national norm. The CMS's empirical Bayes, logistic regression approach to estimate and stabilize the comparisons has generated several criticisms including that it underestimates performance variation, masks the performance of small hospitals (estimates for low-volume hospitals are substantially shrunken toward the national norm), and does not successfully address confounding by hospital-level characteristics. In this context, I identify inferential goals and outline candidate approaches to address the criticisms. These include using a fixed-effects model with hospital-specific intercepts (with the associated wide confidence intervals and high year to year variation for the low volume hospitals), using hospital-level attributes in the risk model or in determining

shrinkage targets, use of a prior distribution other than Gaussian, limiting the shrinkage, replacing the posterior means by ensemble estimates, and modified reporting of results. I outline other features that require consideration, for example selection effects that can bias assessments.

email: *tlouis@jhu.edu*

METHODS FOR PROFILING MEDICAL FACILITIES

John D. Kalbfleisch*, University of Michgan

Kevin Zhi He, University of Michgan

Methods that appropriately account for patient heterogeneity as well as natural unexplained variation among medical facilities are essential in profiling. We develop methods based on models with fixed facility effects that appropriate control for potential confounding between observed patient characterisics and facility effects, and provide more accurate estimates of facility effects that are extreme. Faciltiies are profiled by comparing their outcomes to a national standard based on the outcomes of all similar facilities in the country. The method of profiling is based on an empirical null hypothesis generated by modeling the central part of the distribuiton of observed z statistics corresponding to the facility effects. The approach accounts for facility size and assesses facilities through comparison with those of similar size. This approach provides a robust method for identifying facilities with extreme outcomes. Although based on fixed effects, the method of profiliing can be related in special cases to hierarchical models based on random effects often used in profiling. The methods are illustrated in an example of monitoring outcomes of dialysis facilities in the US.

email: *jdkalbfl@umich.edu*

ON THE ACCURACY OF CLASSIFYING HOSPITALS ON THEIR PERFORMANCE MEASURES

Yulei He*, Centers for Disease Control and Prevention

Sharon-lise Normand, Harvard Medical School

The evaluation, comparison, and public report of health care provider performance is essential to improving the quality of health care. Hospitals, as one type of provider, are often classified into quality tiers (e.g., top or suboptimal) based on their performance data for various purposes. However, potential misclassification might lead to detrimental effects for both consumers and payers. Although such risk has been highlighted by applied health services researchers, a systematic investigation of statistical approaches has been lacking. We assess and compare the expected accuracy of several commonly used classification methods: unadjusted hospital-level averages; shrinkage estimators under a random-effects model accommodating between-hospital variation; and two others based on posterior probabilities. Assuming that performance data follow a classic one-way random-effects model with unequal sample size per hospital, we derive accuracy formulae for these classification approaches and gain insight into how the misclassification might be affected by various factors such as reliability of the data, hospital-level sample size distribution, and cut-off values between quality tiers. The case of binary performance data is also explored using Monte Carlo simulation strategies. We apply the methods to real data and discuss the practical implications.

email: *wdq7@cdc.gov*

31. THE FUTURE OF BIOSTATISTICAL FUNDING MECHANISMS

DISCUSSANTS:

Ciprian M. Crainiceanu, Johns Hopkins University

Francesca Dominici, Harvard University

Debashis Ghosh, Colorado School of Public Health

Lurdes Inoue, University of Washington

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

32. COMPUTER-INTENSIVE BAYESIAN TECHNIQUES AND NEUROSTATISTICS: A PEACEFUL CO-EXISTENCE?

BAYESIAN INFERENCE FOR CLUSTER-STRUCTURED HIGH-DIMENSIONAL ORDINARY DIFFERENTIAL EQUATIONS WITH APPLICATIONS TO BRAIN NETWORKS

Tingting Zhang*, University of Virginia

Brian Caffo, Johns Hopkins University

Qiannan Yin, University of Virginia

Dana Boatman-Reich, Johns Hopkins University

We use ordinary differential equations (ODEs) to model the human brain as a continuous-time dynamic system with biophysical interactions between its components, i.e. brain regions. In contrast to existing ODE models that focus on the connectivity among only a few brain regions, we propose a high-dimensional ODE model for directional connectivity among many brain regions. The new ODE model, called the modular and indicator-based dynamic directional model (MIDDM), features a cluster structure---which consists of several modules of densely connected brain regions, and uses indicators to distinguish significant directional interactions among brain regions from void ones. To perform inference for the MIDDM and also to provide a new statistical approach to quantifying the uncertainty in the ODE model formulation for a complex system, we construct a Bayesian hierarchical model, called Bayesian MIDDM,

for the MIDDM based on basis representation. Specifically, we represent the state functions of the ODE model by cubic spline bases, assign a prior dependent on the ODE model fitting error to basis coefficients, and impose the Potts-model prior on cluster structures and a deliberately designed scaled "spike-and-slab" type of prior on indicators for significant directional effects. The ensuing joint posterior distribution for basis coefficients and the MIDDM parameters has well defined posterior conditional distributions, from which we use a partially collapsed Gibbs Sampler to draw posterior samples. To further speed up the posterior simulation, we employ parallel computing schemes in two Markov Chain Monte Carlo steps. An easy-to-implement hyperparameter selection strategy has also been developed. We apply the proposed Bayesian framework to an auditory electrocorticography dataset to identify significant clusters and directional effects among different brain regions.

email: *tz3b@virginia.edu*

A NOVEL DISTRIBUTIONAL ICA MODEL FOR MULTIMODAL NEUROIMAGING DATA

Ying Guo*, Emory University

Subhadip Pal, Emory University

Jian Kang, University of Michigan

In recent years, the collection of multimodal neuroimaging (e.g. fMRI and DTI) has become common practice in the neuroscience community to advance scientific understanding of brain function and organization. There has been a strong interest in combing different types of imaging because it can capitalize on the complimentary strengths of various modalities. Fusion of multimodal imaging data is a highly challenging problem since data obtain via various modalities have different scales and data representations (scalar/array/matrix). Existing methods usually conduct separate analysis within each modality, which limits their ability to discover multimodal features. In this talk, we present a novel Distributional Independent Component Analysis (D-ICA) framework for decomposing multimodal neuroimaging such as functional MRI (fMRI) and diffusion tensor imaging (DTI). Unlike traditional ICA which separates observed data as a mixture of independent components, the proposed D-ICA represents a fundamentally new approach that aims to perform ICA on the distribution level. The proposed D-ICA method provides a unified framework to extract neural features across imaging modalities that have different scales, representations, signal-to-noise ratios, and intensity. We will discuss the connection and distinction between Standard ICA and D-ICA. Estimation method for the new D-ICA model will be presented. We will illustrate the proposed method through simulation studies and application to a neuroimaging study.

email: *yguo2@emory.edu*

A BAYESIAN GROUP SPARSE MULTI-TASK REGRESSION MODEL FOR IMAGING GENOMICS

Keelin Greenlaw, University of Waterloo

Farouk S. Nathoo*, University of Victoria

Mary Lesperance, University of Victoria

Elena Szefer, Simon Fraser University

Jinko Graham, Simon Fraser University

Recent advances in technology for brain imaging and high-throughput genotyping have motivated studies examining the influence of genetic variation on brain structure. In this setting, high-dimensional regression for multi-SNP association analysis is challenging as the response variables obtained through brain imaging comprise potentially interlinked endophenotypes, and there is a desire to incorporate a biological group structure among SNPs based on their belonging genes. Wang et al. (Bioinformatics, 2012) have recently developed an approach for the analysis of imaging genomic studies based on penalized regression with regularization based on a novel group l_{2,1}-norm penalty which encourages sparsity at the gene level. While incorporating a number of useful features, a shortcoming of the proposed approach is that it only furnishes a point estimate and techniques for obtaining valid standard errors or interval estimates are not provided. We solve this problem by developing a corresponding Bayesian formulation based on a three-level hierarchical model that allows for full posterior inference using Gibbs sampling. Techniques for the selection of tuning parameters are investigated thoroughly and we make comparisons between cross-validation, fully Bayes, and empirical Bayes approaches for the choice of tuning parameters. Our proposed methodology is investigated using simulation studies and is applied to the analysis of a large dataset collected as part of the Alzheimer's Disease Neuroimaging Initiative.

email: *nathoo@math.uvic.ca*

ANALYSIS OF MULTIPLE SCLEROSIS LESIONS VIA A BIVARIATE SPATIAL GLM WITH SPATIALLY VARYING COEFFICIENTS

Timothy D. Johnson*, University of Michigan

Multiple Sclerosis (MS) is a progressive disease in which the myelin sheaths surrounding the axons of the brain and spinal cord are damaged. This leads to dymelination and scarring of the white matter tracks in the brain. There are four main subtypes of MS: 1) relapsing/remitting, 2) secondary progressive, 3) progressive relapsing and 4) primary progressive. Treatment options are subtype dependent and therefore clinicians are interested in using MRI lesion location to predict MS subtype in patients. The data are MS lesions from T1-weighted and T2-weighted MRI images. A bivariate spatial GLM is employed to model the probability of lesion location as a function of subject specific covariates, including disease

subtype, with spatially varying coefficients. Including an a priori probability of MS subtype, we can invert the model to make valid predictions about a specific person's subtype given their imaging data and covariates.

email: *tdjtdj@umich.edu*

33. SURVIVAL ANALYSIS AND GENETICS

USING THRESHOLD REGRESSION TO ANALYZE SURVIVAL DATA FROM COMPLEX SURVEYS: WITH APPLICATION TO NHANES III PHASE II GENETIC DATA

Yan Li, University of Maryland

Dandan Liao, University of Maryland

Mei-Ling Ting Lee*, University of Maryland

In this paper, we propose to extend the threshold regression (TR) model to account for complex sampling designs and to estimate regression parameters of the TR models under complex sampling designs. Innovative features of the proposed method using the pseudo-maximum likelihood estimation technique to estimate the TR model parameters; and proposing computationally-efficient variance estimators that consider the intra-cluster correlation as well as the differential selection probabilities. To demonstrate the usefulness of the pseudo TR models for complex surveys, we present a case example using a complex phase II genetic dataset collected from National Health and Nutrition Examination Survey (NHANES III). Linking the dataset to death certificate records provides an opportunity to use the TR model to analyze time to death due to different types of causes.

email: *mltlee@umd.edu*

EFFICIENT TESTS OF ASSOCIATION FOR SURVIVAL TIMES FROM TWO-PHASE OUTCOME-DEPENDENT SAMPLES

Jerald F. Lawless*, University of Waterloo

In many genetic association studies it is feasible to obtain genomic information on only a fraction of the individuals in a large cohort. In that case it is common to select individuals for genotyping or other genomic measurements according to values of variables that have already been observed; these variables may include actual or surrogate values for responses and covariates of interest. In this talk I consider situations where the response Y is a failure time and we wish to test a null hypothesis of no association between Y and the genomic factors. Efficient likelihood-based score tests will be developed and shown to have a simple common form for a wide class of models and sampling designs. Some comparisons with other methods using weighted estimating functions will be noted. The talk is based in part on joint work with Andriy Derkach and Lei Sun.

email: *jlawless@uwaterloo.ca*

STATISTICAL ISSUES IN GENOME-WIDE ASSOCIATION STUDIES OF BIVARIATE SURVIVAL OUTCOMES

Ying Ding, University of Pittsburgh

Yi Liu, University of Pittsburgh

Qi Yan, University of Pittsburgh

Lars G. Fritsche, University of Michigan

Goncalo G. Abecasis, University of Michigan

Anand Swaroop, National Eye Institute, National Institutes of Health

Emily Y. Chew, National Eye Institute, National Institutes of Health

Daniel E. Weeks, University of Pittsburgh

Wei Chen*, University of Pittsburgh

Age-related Macular Degeneration (AMD) is the leading cause of blindness in the developed world. The genetic causes for disease progression have not been well studied yet. In a National Eye Institute (NEI) funded research project, we aim to identify genetic variants that contribute to disease progression and to build prediction models for clinical guidance. In this talk, we discuss several important statistical issues and challenges in this study. Specifically, to perform genome-wide association studies using eye-level information, we develop a computationally efficient method based on a score test for copula-based bivariate survival model. Using the identified top loci, we establish a prediction model based on semi-parametric Copula to predict progression time for both eyes. Finally, we additionally model their effects on multiple disease progression states through a multistate Markov model, where the transitions among four different AMD states are examined. Both statistical methods and important findings will be presented.

email: *weichen.mich@gmail.com*

GENE-BASED ASSOCIATION ANALYSIS FOR CENSORED TRAITS VIA FIXED EFFECT FUNCTIONAL REGRESSIONS

Ruzong Fan*, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

Yifan Wang, U.S. Food and Drug Administration

Yan Qi, University of Pittsburgh

Ying Ding, University of Pittsburgh

Daniel E. Weeks, University of Pittsburgh

Wei Chen, University of Pittsburgh

Haobo Ren, Regeneron Pharmaceuticals, Inc.

Richard J. Cook, University of Waterloo

Momiao Xiong, University of Texas Health Science Center, Houston

Emily Y. Chew, National Eye Institute, National Institutes of Health

Genetic studies of survival outcomes have been proposed and conducted recently, but statistical methods for identifying genetic

variants that affect disease progression are rarely developed. Motivated by our ongoing real studies, we develop here Cox proportional hazard models using functional regression (FR) to perform gene-based association analysis of survival traits while adjusting for covariates. The proposed Cox models are fixed effect models where the genetic effects of multiple genetic variants are assumed to be fixed. We introduce likelihood ratio test (LRT) statistics to test for associations between the survival traits and multiple genetic variants in a genetic region. Extensive simulation studies demonstrate that the proposed Cox RF LRT statistics have well-controlled type I error rates. To evaluate power, we compare the Cox FR LRT with the previously developed burden test (BT) in a Cox model and sequence kernel association test (SKAT) which is based on mixed effect Cox models. The Cox FR LRT statistics have higher power than or similar power as Cox SKAT LRT except when 50%/50% causal variants had negative/positive effects and all causal variants are rare. In addition, the Cox FR LRT statistics have higher power than Cox BT LRT. The models and related test statistics can be useful in the whole genome and whole exome association studies. An age-related macular degeneration dataset was analyzed as an example.

email: *fanr@mail.nih.gov*

34. MISSING DATA IN NON-INFERIORITY TRIALS

THE IMPACT OF MISSING DATA IN HISTORICAL PLACEBO-CONTROLLED TRIALS

Steven Michael Snapinn*, Amgen Inc.

One unique aspect of non-inferiority trials is that the analysis depends to some extent on the results of the historical placebo-controlled trials used to determine the non-inferiority margin. For this reason, the handling of missing data in both the non-inferiority trial itself and the historical trials impact conclusions. For example, if the control treatment is truly effective, imputing missing values under the null hypothesis in the historical trials will tend to bias the estimate of the treatment effect toward the null, which will result in a smaller margin than if using an imputation method that assumes the data are missing at random. In addition, conceptualizing the evaluation of the experimental treatment as an indirect comparison to placebo can also help inform the appropriate handling of missing data. This presentation will discuss these issues and provide recommendations.

email: *ssnapinn@amgen.com*

MISSING DATA CONSIDERATIONS FOR NON-INFERIORITY TRIALS

Mark D. Rothmann*, U.S. Food and Drug Administration

Missing data considerations for non-inferiority trials will be discussed. Non-inferiority trials have additional aspects to consider on missing data to those of superiority trials. Missing data may not have been addressed in many or some of the studies used to evaluate the effect of the active control. There are also methods (e.g., baseline observation carried forward) for treating missing data that behave like imputation under no treatment difference. While no difference is in the null hypothesis of a superiority comparison, no difference is in the alternative hypothesis of a non-inferiority comparison. Additionally, as there is no difference at baseline between groups in a randomized study, in a repeated measures analysis, an NI margin may apply to the landmark of interest, but not to earlier time points.

email: *mark.rothmann@fda.hhs.gov*

35. IMS MEDALLION LECTURE

MODEL-BASED GEOSTATISTICS FOR PREVALENCE MAPPING IN LOW-RESOURCE SETTINGS

Peter J. Diggle, Ph.D.*, CHICAS, Lancaster University Medical School

In low-resource settings, disease registries do not exist, and prevalence mapping relies on data collected through a finite, often spatially sparse, set of surveys of communities within the region of interest, possibly supplemented by remotely sensed images that can act as proxies for environmental risk factors. A standard geostatistical model for data of this kind is a generalized linear mixed model,

$$Yi \sim \text{Bin}\{mi, P(xi)\} \log[P(xi) / \{1-P(xi)\}] = z(xi)' \beta + S(xi),$$

where Yi is the number of positives in a sample of mi individuals at location xi, z(x) is a vector of spatially referenced explanatory variables and S(x) is a Gaussian process. In this talk, I will first review statistical methods and software associated with this standard model, then consider several methodological extensions whose development has been motivated by the requirements of specific applications. I will focus in particular on prevalence mapping projects that have arisen in connection with pan-African control programs for onchocerciasis (river blindness) and lymphatic filariasis (elephantiasis). These vectorborne diseases are major public health problem in the wet tropical regions of the world, including most of sub-Saharan Africa. Multi-national control programs using mass administration of a protective drug, Mecitzan, have been very successful, with more than 60 million treatments to date over 19 countries. However, the programs has been hampered by the recognition that people heavily infected with a third disease, Loa loa (eyeworm) parasite, are at risk of severe, occasionally fatal, adverse reaction to Mectizan. Before the drug is administered in

a community, it is relatively easy to estimate the prevalence of eyeworm infection, harder (and more expensive) under field conditions to estimate how many people are "heavily infected," one definition of which is that they as carrying more than 8,000 parasites per ml of blood. To address this problem we develop a joint model for communitylevel prevalence and the proportion of highly infected individuals in the community.

email: *p.ediggle@lancaster.ac.uk*

36. ANALYSIS OF IMAGING DATA

MIXED EFFECTS MODELS TO FIND DIFFERENCES IN MULTI-SUBJECT FUNCTIONAL CONNECTIVITY

Manjari Narayan*, Rice University

Genevera I. Allen, Rice University

Many complex brain disorders such as autism spectrum disorders exhibit a wide range of symptoms and disability. To understand how brain communication is impaired in such conditions, functional connectivity studies seek to understand individual differences in brain network structure in terms of covariates that measure symptom severity. In practice, however, functional connectivity is not observed but estimated from complex and noisy neural activity measurements. Imperfect subject network estimates can compromise subsequent efforts to detect covariate effects on network structure. We address this problem in the case of Gaussian graphical models of functional connectivity, by proposing novel two-level models that treat both subject level networks and population level covariate effects as unknown parameters. To account for imperfectly estimated subject level networks when fitting these models, we propose two related approaches R2 & R3 based on resampling, random adaptive penalization and random effects test statistics. Simulation studies using realistic graph structures reveal that R2 and R3 have superior statistical power to detect covariate effects compared to existing approaches, particularly when the number of within subject observations is comparable to the size of subject networks. Using our novel models and methods to study parts of the ABIDE dataset, we find evidence of hypoconnectivity associated with symptom severity in Autism spectrum disorders, in frontoparietal and limbic systems as well as in anterior and posterior cingulate cortices.

email: *manjari@rice.edu*

DEFORMATION ANALYSIS OF DIFFUSION TENSOR DATA USING RANDOM FORESTS

Neda Sadeghi*, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

M. Okan Irfanoglu, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health and Henry M. Jackson Foundation

Amritha Nayak, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health and Henry M. Jackson Foundation

Cibu Thomas, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health and Center for Neuroscience and Regenerative Medicine

Carlo Pierpaoli, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

The standard practice for analyzing neuroimaging data (e.g. diffusion tensor imaging) is to register the individual images to a common coordinate space and perform a voxel-wise comparison of diffusion derived metrics between groups. Neuroimaging data is inherently high dimensional and massive multiple comparisons correction is necessary in the voxel-wise analysis; the choice of the method used for correction can produce differing results. Another important method, that is less common in diffusion tensor imaging, is to analyze the deformation fields that map individual images from their native space to a common template. In this work we use the log of the Jacobian of the deformation fields in regions of interest as input to random forests. Random forests are a set of data driven classification algorithms that highlight important features. A random forest is iteratively fit to the data, at each iteration the least important variables are eliminated until the current out of bag (OOB) error becomes larger than the previous OOB rate + OOB standard error. This results in regions that achieve high classification rate and are important in characterizing the disease. We present the methodology applied to a group of patients that have brain atrophy.

email: *neda.sadeghi@nih.gov*

ON ESTIMATING FUNCTIONAL CONNECTIVITY FOR NEUROIMAGING DATA

Ivor Cribben*, University of Alberta

We discuss functional connectivity analyses of the human brain. In the first part of this talk, we compare several sparse graphical estimation procedures (graphical lasso, SCAD, DP-glasso, …) and several selection criteria (AIC, BIC, CV, …) using both simulated multivariate normal data and autocorrelated data. We use evaluation criteria to compare the models and thoroughly discuss the superiority and deficiency of each of them. We estimate the functional connectivity networks between regions of interest (ROIs) from a resting state functional magnetic resonance imaging (fMRI) experiment and from a language processing experiment using the various

procedures. In the second part of the talk, we propose a flexible general linear framework for estimating functional connectivity that accounts for 1) temporal autocorrelation in a non-parametric manner and 2) heterogeneity across subjects by allowing subject-specific estimates of the variance. We carry out a simulation study to assess the performance of our proposed method with respect to Type I and II errors, and we illustrate the utility of this new method on an fMRI study of depression.

email: *cribben@ualberta.ca*

MULTILINEAR PRINCIPAL COMPONENTS ANALYSIS IN SPATIALLY VARYING COEFFICIENT MODEL FOR NEUROIMAGE DATA

Tianming Zhang*, University of South Carolina

Yanyuan Ma, University of South Carolina

Linglong Kong, University of Alberta

This paper is inspired by spatially varying coefficient model (SVCM) for neuroimaging data with jump discontinuities (Zhu, Fan and Kong 2013). SVCM is developed to described the varying association among brain image measures. We aim to find a more effective and accurate way to estimate the individual image variations and its covariance function in SVCM by applying multilinear functional principal components analysis (MPCA). The core of our method is deducting the dimension of image matrices (2D or 3D) by applying generalized low rank approximation of matrices. Our dimension deduction method keeps the natural matrix structure of measurements without vectorizing matrices of image data. And we simplify the traditional propagation-separation method by using different bandwidth selection methods to reduce computation time without losing accuracy. The simulation and real data analysis comfirm our improvments.

email: *Zhang373@email.sc.edu*

ASSESSING UNCERTAINTY IN DYNAMIC FUNCTIONAL CONNECTIVITY

Maria Aleksandra Kudela*, Indiana University School of Public Health

Jaroslaw Harezlak, Indiana University School of Public Health

Martin A. Lindquist, John Hopkins Bloomberg School of Public Health

Functional connectivity (FC), the study of the statistical association between anatomically distinct time-series (Friston 1994, 2011), has become one of the primary areas of research in the field surrounding fMRI (functional magnetic resonance imaging). While, for many years researchers implicitly assumed that FC was stationary across time; it has recently become increasingly clear that the ability to assess dynamic changes in FC is critical for better understanding of the inner workings of the human brain (Hutchison et al. 2014). Currently, the most common strategy to estimate these dynamic changes is by applying the sliding window technique. However, possibly its greatest shortcoming is the inherent variation present in the estimate, even for null data, which is easily confused with true time-varying changes in connectivity (Lindquist et al. 2014). This can have serious consequences as even spurious fluctuations caused by noise can easily be confused with important signal. For these reasons, assessment of uncertainty in the sliding window correlation estimates is of critical importance. Here we propose a new approach that combines the MLPB and sliding-window techniques, to assess the uncertainty in dynamic FC estimates by providing its confidence bands. Both numerical results and an application to fMRI study are presented.

email: *maria.kudela@gmail.com*

MODELING CONNECTIVITY IN HIGH-DIMENSIONAL BRAIN SIGNALS

Yuxiao Wang*, University of California, Irvine

Chee-Ming Ting, Universiti Teknologi Malaysia

Hernando Ombao, University of California, Irvine

We develop a novel approach to modeling connectivity in high dimensional brain signals. In the approach, we model the cortical activity using linear mixture of latent factor activities that follows a vector autoregressive (VAR) process. The frequency-specific connectivity on the cortical surface can be characterized by the latent factor activity and its loading matrix. The primary motivation for this work is modeling connectivity among regions on the cortical surface using multi-channel scalp electroencephalograms (EEG). Modeling connectivity between brain regions is difficult under high dimensionality of the anatomical parcellation on the cortical surface. We present a modeling procedure that addresses a number of challenges in high dimensional brain signals. In the first step, we estimate the sources using imaging method with anatomical constraints. In the second step, to estimate temporal dependency between regions on the cortex, we fit a latent process with a vector autoregressive (VAR) structure. The VAR parameters are then estimated to produce measurements of the cortical connectivity. The potential utility of the proposed approach is demonstrated in the analysis of resting-state EEG data.

email: *yxwang87@gmail.com*

STATISTICAL APPROACHES FOR EXPLORING BRAIN CONNECTIVITY WITH MULTI-MODAL NEUROIMAGING DATA

Phebe B. Kemmer*, Emory University

Ying Guo, Emory University

DuBois Bowman, Columbia University

Functional connectivity (FC) is often the main objective of fMRI studies using data-driven methods such independent component analysis (ICA). With the advent of diffusion tensor imaging (DTI) and probabilistic tractography we can also evaluate structural connectivity (SC) in the brain. Despite the immense research devoted to FC, important questions are beginning to emerge. For example, do structural connections (SC) underlie functionally connected brain regions? Do structural connections underlying an FC network differ between subpopulations? Is knowledge about SC helpful in informing the reliability of FC results? To address these questions, we develop a new measure to quantify the strength of structural connectivity (sSC) underlying FC networks, and derive test both for its statistical significance and for subgroup comparisons (e.g. healthy vs depressed patients). We also demonstrate that sSC is associated with the reproducibility of identified FC networks, such that FC networks with strong underlying SC tend to be more reliable. We demonstrate the performance of our measure with simulation studies and illustrate the method with an application to a resting-state fMRI and DTI study of major depressive disorder (MDD).

email: *brennep@gmail.com*

37. BAYESIAN CLINICAL TRIALS

CREDIBLE SUBGROUP INFERENCE FOR BOUNDING THE BENEFITING SUBPOPULATION FOR MANY TREATMENTS AND MULTIPLE ENDPOINTS

Patrick Schnell*, University of Minnesota

Qi Tang, AbbVie

Peter Mueller, University of Texas, Austin

Bradley P. Carlin, University of Minnesota

Many new experimental treatments outperform the current standard only on a subset of the population. The credible subgroups method provides a pair of bounding subgroups for the benefiting subgroup constructed so that one contains only patients with an expected benefit and the other contains all patients with an expected benefit. However, when more than two treatments and multiple endpoints are under consideration, there are many possible requirements for a particular treatment to be beneficial. We extend the credible subgroups method to handle such cases, and apply the extended method to an example dataset from an Alzheime's Disease treatment trial.

email: *schn0956@umn.edu*

INCORPORATION OF STOCHASTIC ENGINEERING MODELS AS PRIOR INFORMATION IN BAYESIAN MEDICAL DEVICE TRIALS

Rajesh Nair*, U.S. Food and Drug Administration

Tarek Haddad, Medtronic

Adam Himes, Medtronic

Laura Thompson, U.S. Food and Drug Administration

Telba Irony, U.S. Food and Drug Administration

Stochastic engineering models are being increasingly used during the product development process for medical devices. These models have the capability to simulate virtual patient outcomes. Incorporation of these models as prior knowledge in a Bayesian clinical trial design can provide benefits of decreased sample size and trial length while still controlling type I and type II error rates. This paper presents a method for augmenting a clinical trial using virtual patient data, where the number of virtual patients is based on the similarity between modeled and observed data. The use of this method is illustrated by a case study based on a model for cardiac lead fracture.

e-mail: *rajesh.nair@fda.hhs.gov*

BAYESIAN ADAPTIVE DOSE FINDING FOR COMBINATION THERAPY IN PHASE I ONCOLOGY TRIALS

Chenyi Pan*, University of Virginia

Yun Shen, Bristol-Myers Squibb

Helen Zhou, Bristol-Myers Squibb

Parul Gulati, Bristol-Myers Squibb

Xiaowei Guan, Bristol-Myers Squibb

Katy Simonsen, Bristol-Myers Squibb

WITHDRAWN

Treating patients with a combination of drugs is becoming commonplace in oncology trials. A new Bayesian adaptive approach to find the maximum-tolerated dose (MTD) in phase I oncology trials for combination therapy is proposed. The approach relies on the joint toxicity model proposed in this paper, i.e. BLRM-Copula model. The proposed design allows to incorporate the uncertainties in single agent toxicity profile and supports dose escalation decision making from a Bayesian model-based perspective. The modeling and decision making components investigated here are flexible enough to be extended to more complex settings. The operating characteristics of the BLRM-Copula model are accessed through the simulation study by comparing with the Bayesian logistic regression model (BLRM). The comparisons revealed comparable performance of these two models while obvious advantage in terms of sample size when applying BLRM-Copula model. The design and BLRM-Copula model discussed in this paper has already been implemented in oncology combination trials to identify MTD.

e-mail: *cp2xd@virginia.edu*

USING DATA AUGMENTATION TO FACILITATE CONDUCT OF PHASE I/II CLINICAL TRIALS WITH DELAYED OUTCOMES

Ick Hoon Jin*, University of Notre Dame
Suyu Liu, University of Texas MD Anderson Cancer Center
Peter F. Thall, University of Texas MD Anderson Cancer Center
Ying Yuan, University of Texas MD Anderson Cancer Center

A practical impediment in adaptive clinical trials is that outcomes must be observed soon enough to apply decision rules to choose treatments for new patients. For example, if outcomes take up to six weeks to evaluate and the accrual rate is one patient per week, on average three new patients will be accrued while waiting to evaluate the outcomes of the previous three patients. The question is how to treat the new patients. This logistical problem persists throughout the trial. Various ad hoc practical solutions are used, none entirely satisfactory. We focus on this problem in phase I–II clinical trials that use binary toxicity and efficacy, defined in terms of event times, to choose doses adaptively for successive cohorts. We propose a general approach to this problem that treats late-onset outcomes as missing data, uses data augmentation to impute missing outcomes from posterior predictive distributions computed from partial follow-up times and complete outcome data, and applies the design's decision rules using the completed data. We illustrate the method with two cancer trials conducted using a phase I–II design based on efficacy–toxicity trade-offs, including a computer stimulation study.

e-mail: *ijin@nd.edu*

CONTROL CHARTS FOR MONITORING ACCUMULATING ADVERSE EVENT COUNT FREQUENCIES FROM SINGLE AND MULTIPLE BLINDED TRIALS

A. Lawrence Gould*, Merck Research Laboratories

Monitoring accumulating information about drug safety in terms of the numbers of adverse events reported from trials in a drug development program is conventional practice. Estimates of between-treatment adverse event risk differences can be obtained readily from unblinded trials with differences among trials accounted for using conventional statistical methods. Recent regulatory guidelines require monitoring the cumulative frequency of adverse event reports to identify possible between-treatment adverse event risk differences without unblinding ongoing trials. Conventional statistical methods for assessing assessing between-treatment adverse event risks cannot be applied when the trials are blinded. However, CUSUM charts can be used to monitor the accumulation of adverse event occurrences on an ongoing basis. CUSUM charts for monitoring adverse event occurrence are based on assumptions about the process generating the adverse event counts in a trial as expressed by informative prior distributions. We describe the construction of control charts for monitoring adverse event occurrence based on statistical models for the processes, characterize their statistical properties, and describe how to construct useful prior distributions. Application of the approach to two adverse events of interest in a real trial gave nearly identical results for binomial and Poisson observed event count likelihoods

email: *goulda@merck.com*

APPLICATION OF BAYESIAN METHODS FOR MAKING GO/NO-GO DECISION IN CLINICAL TRIALS WITH AN EXAMPLE

Rodney Croos-Dabrera*, Astellas Pharma Development
Misun Lee, Astellas Pharma Development

Planning late stage clinical developments such as Phase III is costly and time consuming. Having appropriate statistical tools for making an informed and quantitative Go/No-Go decision at the end of the proof-of-concept (PoC) study is essential for clinical development of an experimental drug. In the development of Bayesian methods for Go/No-Go decision in clinical trials, a significant progress has been made in recent years. In this talk, we will discuss how such method can be used to make an informed Go/No-Go decision using a real data example in Infectious Disease therapeutic area. In particular, we will discuss modified approaches to this existing methodology that may fit better for rare disease patients. For this illustration, one of the main drivers would be the probability of technical success (PTS), that is, a probability of success in both PoC and Phase III published by Michael Hayet. al. (2014). Additionally, we will discuss how to utilize graphical summary display of PTS for rare diseases.

email: *rodney.dabrera@gmail.com*

38. DIAGNOSTIC AND SCREENING TESTS

COMPARING PAIRED DIAGNOSTIC TESTS BASED ON JOINT TESTING OF THE AUC AND THE YOUDEN INDEX

Jingjing Yin*, Georgia Southern University
Lili Tian, University at Buffalo
Hani Samawi, Georgia Southern University

In the ROC analysis, the area under the ROC curve (AUC) serves as an overall measure of a biomarker/diagnostic test's accuracy. Another popular index is Youden index (J), which corresponds to the maximum sum of sensitivity and specificity thus can be used for diagnostic threshold optimization. Although researchers mainly evaluate the diagnostic accuracy using the AUC, for the purpose of making diagnosis, Youden index provides a direct measure of the diagnostic accuracy at the optimal threshold and hence should

be taken into consideration in addition to the AUC. Our previous research proposed the joint confidence region of AUC and Youden index for a single test. Furthermore, it is very common to compare the diagnostic accuracy of two correlated tests and see if one biomarker is more preferable in terms of both summary indices. This can be done by testing Ho: AUC1-AUC2=0 and J1-J2=0 versus Ha: AUC1-AUC2>0 and J1-J2>0. The existing approach for testing such order restrictive hypothesis is the intersection-union test (IUT), which marginally test the AUC and the Youden index independently. We propose an alternative test procedure in both parametric and non-parametric settings, which is shown by simulations to be much more powerful than IUT test under the alternative and maintain the type I error rate under the null.

email: *jyin@georgiasouthern.edu*

MODELING AGREEMENT BETWEEN MANY RATERS USING AN ORDERED CLASSIFICATION SCALE

Kerrie P. Nelson*, Boston University
Don Edwards, University of South Carolina

Ordinal categorical scales are commonly used in screening and diagnostic tests to classify a patient's disease status. However, severe discrepancies between different raters' classifications in common diagnostic procedures have motivated large-scale studies to be conducted incorporating the classifications of multiple raters to assess accuracy and agreement. Limited methods are available to model the agreement between many raters in a unified comprehensive manner. In this talk we describe a flexible model-based approach and measure of agreement based upon the class of generalized linear mixed models that can be used to assess agreement between large numbers of raters. We apply the methods to a recent large-scale cancer agreement study.

email: *kerrie@bu.edu*

ON THE USE OF MIN-MAX COMBINATION OF BIOMARKERS TO MAXIMIZE THE PARTIAL AREA UNDER THE ROC CURVE

Hua Ma*, Duke University
Susan Halabi, Duke University

The partial area under the curve (pAUC) is an important summary index focusing on the range of practical/clinical relevance in the Receiver Operating Characteristic (ROC) curve analysis. When multiple continuous-scaled biomarkers are available, finding optimal linear combination to maximize pAUC is challenging. We proposed to extend the min-max method to the estimation of pAUC and compared its performances with different existing methods. Simulations were conducted to investigate the performance of different methods based on their abilities to yield the largest pAUC estimates. Different distributions of biomarker values, shapes of ROC curves, false positive fraction ranges, and sample size configurations were considered. Mean and standard deviation of pAUC estimates through resubstitution and leave-one-out cross validation were obtained. Our results demonstrate that the proposed method provides larger pAUC estimates under the following two important practical scenarios: (1) multivariate data for non-diseased and diseased subjects have unequal variance-covariance matrices and ROC curves generated from individual biomarker are relative close regardless of the latent normality distributional assumption; or (2) ROC curves generated from individual biomarker have straight line shapes. In conclusion, the proposed method seems robust and may be used in the estimation of pAUC in many practical situations.

email: *hua.ma@duke.edu*

ESTIMATION OF DISCRETE SURVIVAL FUNCTION THROUGH THE MODELING OF DIAGNOSTIC ACCURACY FOR MISMEASURED OUTCOME DATA

Abidemi K. Adeniji*, Boehringer Ingelheim Pharmaceuticals
Hee-Koung Joeng, University of Connecticut
Naitee Ting, Boehringer Ingelheim Pharmaceuticals
Ming-Hui Chen, University of Connecticut

Standard survival methods are inappropriate for mismeasured outcomes. Previous research has shown that outcome misclassification can bias estimation of the survival function. We develop methods to accurately estimate the survival function when the diagnostic tool used to measure the outcome of disease is not perfectly sensitive and specific. Since the diagnostic tool used to measure disease outcome is not the gold standard, the true or error-free outcomes are latent. Our method uses the negative predictive value (NPV) and the positive predictive values (PPV) of the diagnostic tool to construct a bridge between the error-prone outcomes and the true outcomes. We formulate an exact relationship between the true (latent) survival function and the observed (error-prone) survival function as a formulation of time-varying NPV and PPV. We specify models for the NPV and PPV that depend only on parameters that can be easily estimated from a fraction of the observed data. Furthermore, we conduct an in depth study to accurately estimate the latent survival function based on the assumption that the biology that underlies the disease process follows a stochastic process. We further examine the performance of our method by applying it to the VIRAHEP-C data.

email: *abidemi.adeniji@gmail.com*

THE OPTIMAL LENGTH OF A SEQUENCE OF TESTS FOR CLASSIFICATION TASKS

Christine M. Schubert Kabban, Air Force Institute of Technology

Donna K. McClish, Virginia Commonwealth University

Combining classification systems in order to improve diagnostic accuracy is by no means a new concept. However, emphasis on reducing the cost of testing has direct implication for any combination of tests. Cost here, refers to the expense of operating a combination of tests, whether that expense be measured financially in dollars, in the time required to complete the series of tests, or by some other criteria. This work intends to describe a means by which researchers may be able to determine the sequence of tests that provides optimal performance (accuracy) for classification while maintaining minimal operational cost for tests with two outcomes. This is accomplished by using tolerances on a newly constructed, weighted function of sequence accuracy and operational cost in order to determine the best sequence of tests for the diagnostic task. Computational formulas are presented for sequence accuracy, as represented by the probabilities of true and false positive, as well as for sequence cost. Weighting is provided so that trade-offs between accuracy and operational cost, as well as misclassification costs and class prevalences may be made. Simulated results demonstrate the effects of individual test accuracy and operational cost, prevalence, misclassification costs, and correlation on the optimal sequence.

email: *christine.schubertkabban@afit.edu*

A PLACEMENT VALUE BASED APPROACH TO CORRELATED AND CONCAVE ROC CURVES WITH ORDER CONSTRAINTS

Zhen Chen*, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

Sung Duk Kim, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

Beom Seuk Hwang, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

In many diagnostic accuracy studies with multiple correlated tests for detecting a disorder, it is possible that the tests are ordered in their performance a priori. When these prior information are available, it is important to incorporate them in the estimation of the associated ROC curves, as this can potentially improve statistical efficiency. Further, it is often desirable to consider ROC curves that are concave, a feature that are consistent with optimal decision principles. Motivated by these considerations, we propose a new approach to estimating multiple correlated ROC curves using placement values. The concavity constraint is constructed by recasting the random variable for the placement values as a product of a uniform and another arbitrary random variables, while the ordering constraints are achieved through product measures in the context of mixture distributions. To allow flexible distributions for the test scores and for the derived placement values, we use Dirichlet process mixture priors. Through simulation studies, we demonstrate that the proposed approach has good performance. We illustrate the methodology with an application to the Physician Reliability Study that investigated the diagnosis of endometriosis using different combinations of clinical information.

email: *chenzhe@mail.nih.gov*

AN APPLICATION OF FACTOR ANALYSIS IN DEVELOPING AN ABBREVIATED QUESTIONNAIRE: CASE STUDY FROM NEUROLOGY

Jaya**want Mandrekar***, Mayo Clinic

Many clinical studies collect data using several survey instruments. These survey instruments may be filled out either by patients or care givers. Survey instruments that contain too many questions take longer time to administer, maybe burdensome for sicker patients and also problematic if scoring algorithm is complex. Longer surveys increase likelihood of missing or less reliable information if responses are not captured adequately. This can lead to biased conclusions. A scientifically selected smaller subset of the original survey instrument as an alternative may require less time to complete and may capture necessary information more reliably. Factor analysis can be one of many statistical tools that can allow us to reduce the dimensionality i.e. select a smaller subset of questions. This apporach also aids in identification of internally consistent question domains. A real life application from neurology research where a much smaller questionnaire was developed using a combination of statistical techniques and clinical insights will be presented.

email: *mandrekar.jay@mayo.edu*

39. IMS MEDALLION LECTURE

IMPROVING POWER WITH GENERALIZED ESTIMATING EQUATIONS IN SMALL-SAMPLE LONGITUDINAL STUDY SETTINGS

Philip M. Westgate*, University of Kentucky

Woodrow W. Burchett, University of Kentucky

Generalized estimating equations (GEE) are often used for the marginal analysis of longitudinal data. Although much work has been done to improve the validity of GEE for the analysis of data arising from small-sample studies, little attention has been given to power in such settings. Therefore, we propose a valid GEE approach to improve power in small-sample longitudinal study settings. Specifically, we use a modified empirical sandwich covariance matrix

estimator within correlation structure selection criteria and test statistics. Use of this estimator can improve the accuracy of selection criteria and increase the degrees of freedom to be used for inference. Resulting power increases will be shown via a simulation study and application example.

email: *philip.westgate@uky.edu*

ON THE BRIDGE BETWEEN BRIDGE DISTRIBUTIONS, MARGINALIZED

Geert Molenberghs*, Universiteit Hasselt and Katholieke Universiteit Leuven, Belgium

The generalized linear mixed model is commonly used for the analysis of hierarchical non-Gaussian data. It combines an exponential family model formulation with normally distributed random effects. A drawback is the difficulty of deriving convenient marginal mean functions with straightforward parametric interpretations. Several solutions have been proposed, including the marginalized multilevel model (directly formulating the marginal mean, together with a hierarchical association structure) and the bridging approach (choosing the random-effects distribution such that marginal and hierarchical mean functions share functional forms). Another approach, useful in both a Bayesian and a maximum likelihood setting, is to choose a random-effects distribution that is conjugate to the outcome distribution. In this paper, we contrast the bridging and conjugate approaches. For binary outcomes, using characteristic functions and cumulant generating functions, it is shown that the bridge distribution is unique. Self-bridging is introduced as the situation in which the outcome and random-effects distributions are the same. It is shown that only the Gaussian and degenerate distributions have well-defined cumulant generating functions for which self-bridging holds.

email: *geert.molenberghs@uhasselt.be*

A COMPARISON OF THREE MODELS IN MULTIVARIATE BINARY LONGITUDINAL ANALYSIS

Hissah Alzahrani*, Florida State University

Elizabeth Slate, Florida State University

Multivariate longitudinal data analysis plays an important role in many biomedical and social problems. In this article, we present three methods for analyzing multiple and correlated binary outcomes; each one can be beneficial for determined aims. We review method one and method two and we proposed method three. The three methods estimate the marginal means using the GEE approach for multivariate binary longitudinal data. The first method addresses the question of estimating one group of covariate parameters for many binary outcomes while accounting for their multivariate structure. The second method addresses the question of estimating the covariate parameters for each binary outcome separately. The third method is an estimation of the covariate parameters for each combination of outcomes. Our goal is to investigate the difference among the parameter estimations of the three methods. In the simulation element, we present many scenarios related to different correlation structures. In the application element, we present a follow up study (Florida Dental Care Study) that measured three binary outcomes and five covariates in four intervals. That particular study is a useful explanation of the variation between outcomes since the outcomes were highly correlated.

email: *hahzahrani@gmail.com*

DISCREPANCY-BASED PARAMETER ESTIMATION FOR BALANCING EFFICIENCY AND ROBUSTNESS IN FITTING STATE-SPACE MODELS

Nan Hu*, University of Iowa

Joseph Cavanaugh, University of Iowa

In the state-space modeling framework, parameter estimation is often accomplished by maximizing the innovations Gaussian log-likelihood. The maximum likelihood estimator (MLE) is efficient when the normality assumption is satisfied. However, in the presence of contamination, the MLE suffers from a lack of robustness. Basu, Harris, Hjort, and Jones (1998) introduced a discrepancy measure (BHHJ) with a nonnegative tuning parameter that controls the trade-off between robustness and efficiency. In this talk, we propose a new parameter estimation procedure based on the BHHJ discrepancy for fitting state-space models. As the tuning parameter is increased, the estimation procedure becomes more robust but less efficient. We investigate the performance of the procedure in a comprehensive simulation study, and illustrate its utility in a practical application. In addition, we provide guidelines on how to choose an appropriate tuning parameter.

email: *nan-hu@uiowa.edu*

IMPROVED POWER IN CROSSOVER DESIGNS THROUGH LINEAR COMBINATIONS OF BASELINES

Thomas Jemielita*, University of Pennsylvania

Mary Putt, University of Pennsylvania

Devan Mehrotra, Merck Research Laboratories

In a crossover design with continuous outcomes (e.g., blood pressure), baseline and post-baseline responses are obtained in each treatment period. The baselines can be utilized as covariate(s) in an analysis of covariance (ANCOVA) to increase the precision of the treatment effect estimate. Previous authors have noted that the potential efficiency gain from using baselines depends on the joint covariance structure of all the baseline and post-baseline responses. We show how the underlying covariance structure can be leveraged to find an optimal linear combination of the baselines

so as to minimize the theoretical variance of the ANCOVA-based estimated treatment effect. We do this for balanced 2x2, 3x3, and 4x4 crossovers under four commonly seen covariance structures. We also develop an adaptive method in which first a suitable covariance structure for the given dataset is selected via AICC values, and then the corresponding optimal baseline covariate combination is used in the ANCOVA. We show that, relative to previously published methods, the proposed method leads to sizable gains in power, while maintaining the nominal type I error rate.

email: *thomasjemielita@gmail.com*

A CAUTIONARY NOTE ON USING GENERALIZED ESTIMATING EQUATIONS TO ESTIMATE TRANSITION MODELS

Joe D. Bible*, 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

Danping Liu, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

Generalized estimating equations (GEE) are commonly used to estimate transition models, where independent working correlation is a convenient choice in practice. It is often ignored that such GEE models lack robustness to different choices of the working correlation. We assume that the true model is from a wide class of random effects models, where the random effects dictate the correlation structure of multiple transitions from the same subject. In most situations, the parameter of interest is the population-average transition probabilities where each subject in the population should contribute equally to this average. However, GEE with independent correlation treats all the transitions equally, and as a result, subjects with more 0-1 transitions get over-weighted, and the estimated transition probabilities are biased. Through asymptotic bias calculations and finite sample simulations, we demonstrate that GEE with unstructured correlation provides accurate estimators, because the estimating equations can be viewed as quasi-score equations. We advocate the specification of correct working correlation when fitting marginal transition models.

email: *jbible831@gmail.com*

SOME STRUCTURED ANTEDEPENDENCE MODELS FOR MULTIVARIATE LONGITUDINAL DATA

Chulmin Kim*, University of West Georgia

In many of medical and biological studies, two or more attributes are measured on each subject over the repeated time, yielding-multivariate longitudinal data. Antedependence (AD) models which are generalization of Autoregressive (AR) models that allow the variances and same-lag correlations to vary over time can be useful for constructing covariance for longitudinal data. Zimmerman(1997) introduced a structured AD model which is more useful than unstructured AD models for some non-stationary longitudinal data. We generalize the univariate AD model to multivariate AD model (MAD). We would like to construct some structured MAD models for the covariance structure ofmultivariate longitudinal dataand illustrate the properties of the structured MAD.

email: *ckim@westga.edu*

40. ORAL POSTERS: MACHINE LEARNING

40a. INVITED ORAL POSTER: REGRESSION FOR BLOCK-MISSING MULTI-MODALITY DATA

Guan Yu*, University of North Carolina, Chapel Hill

Quefeng Li, University of North Carolina, Chapel Hill

Yufeng Liu*, University of North Carolina, Chapel Hill

In modern scientific research, many data are collected from multiple modalities (sources or types). Since different modalities could provide complementary information, sparse regression methods using multi-modality data have the potential to deliver better prediction performance. However, one special challenge for using multi-modality data is the challenging of missing data. In practice, the observations of a certain modality can be missing completely, i.e., a complete block of the data is missing. In this work, we propose a new two-step sparse regression method for block-missing multi-modality data. Rather than deleting samples with missing data or imputing the missing observations, the proposed method makes use of all available information without any imputation. The effectiveness of the proposed method is demonstrated by theoretical studies, simulated examples, and a real data example from the Alzheimer's Disease Neuroimaging Initiative.

email: *yfliu@email.unc.edu*

40b. INVITED ORAL POSTER: A NOVEL AND EFFICIENT ALGORITHM FOR DE NOVO DISCOVERY OF MUTATED DRIVER PATHWAYS IN CANCER

Binghui Liu, Northeast Normal University, China

Xiaotong Shen, University of Minnesota

Wei Pan*, University of Minnesota

Next-generation sequencing studies on cancer somatic mutations have discovered that driver mutations tend to appear in most tumor samples, but they barely overlap in any single tumor sample, presumably because a single driver mutation can perturb the whole pathway. Based on the corresponding new concepts

of coverage and mutual exclusivity, new methods can be designed for de novo discovery of mutated driver pathways in cancer. Since the computational problem is a combinatorial optimization with an objective function involving a discontinuous indicator function in high dimension, many existing optimization algorithms, such as a brute force enumeration, gradient descent and Newton's methods, are not practically feasible or directly applicable. We develop a new algorithm based on a novel formulation of the problem as non-convex programming and non-convex regularization. The method is computationally more efficient, effective and scalable than existing Monte Carlo searching and several other algorithms, which have been applied to The Cancer Genome Atlas (TCGA) project. We demonstrate its promising performance with application to two cancer datasets to discover de novo mutated driver pathways.

email: *weip@biostat.umn.edu*

40c. EXTENDING THE METHOD, FEATURE AUGMENTATION VIA NONPARAMETRICS AND SELECTION, TO THE ORDINAL RESPONSE SETTING

Kyle L. Ferber*, Virginia Commonwealth University

Kellie J. Archer, Virginia Commonwealth University

Feature Augmentation via Nonparametrics and Selection (FANS) is a binary classification procedure that has shown promising results in high-dimensional learning problems. Instead of including the original form of the predictors in the model, FANS fits an additive model of augmented features. An augmented feature is defined as the log ratio of the conditional marginal density estimates of the two classes for a given predictor. In this work, we extended FANS to the ordinal response setting in which there are K> 2 ordered outcome classes. We defined K- 1 augmented features for each original predictor and developed a model fitting algorithm that utilizes data splitting and prediction averaging to make efficient use of the data. Our method will enable researchers to develop a model with high predictive accuracy and extract a parsimonious subset of the high-dimensional feature set that is jointly predictive of the ordinal outcome. We present the results of our analysis of a high-dimensional gene expression dataset to demonstrate the method's performance.

email: *ferberkl@vcu.edu*

40d. PENALIZED BAYESIAN CUMULATIVE LOGIT MODEL FOR HIGH-DIMENSIONAL DATASETS

Qing Zhou*, Virginia Commonwealth University

Kellie J. Archer, Virginia Commonwealth University

Motivated by the Bayesian LASSO proposed by Park and Casella (2008), we developed an ordinal response model that incorporates a penalty term so that a parsimonious model can be obtained. Our penalized ordinal Bayesian method includes the likelihood of the cumulative logit model combined with a Laplace prior (e.g. double exponential prior) where the penalization parameter is chosen by imposing different gamma priors. We will illustrate the utility of this method using both simulated data and a high-throughput genomic dataset. In our simulation study, we compare our penalized ordinal Bayesian model using different priors to a penalized cumulative logit model using a frequentist approach (generalized monotone incremental forward stage-wise method) in term of their abilities to predict the ordinal response and to correctly incorporate true predictors from noise predictors into the model. We also demonstrate application of our method to predict stage of liver disease (normal, cirrhotic but without hepatocellular carcinoma, hepatocellular carcinoma) using methylation data assayed by the Illumina GoldenGate Methylation BeadArray Cancer Panel I.

email: *zhouq3@vcu.edu*

40e. SPARSE MEDIATION ANALYSIS FOR HIGH-DIMENSIONAL MEDIATORS

Yi Zhao*, Brown University

Xi Luo, Brown University

In empirical research, scientists are interested in testing causal mechanisms through which a treatment affects the outcomes. For cases with multiple causally dependent mediators, the existing methods require fitting SEMs with all the mediators as predictors, which are not applicable to the problem of high-dimensional mediators. In this study, instead of specifying the causal relationships between the mediators, we propose an alternative representation of the causal mechanisms with large number of mediators under SEM framework. An l1-regularization on the mediation effects, which are represented by the product of model coefficients, is introduced to select the causal mechanisms and estimate mediation effects simultaneously. A novel penalty function is proposed to make the objective function convex and computationally feasible. To estimate the parameters, an ADMM combined with augmented Lagrangian is proposed. For the non-smooth part of the objective function, the solution can be solved in explicit forms. Compared to the marginal mediation approach with multiple-testing adjustment, our method can attain higher area under the ROC curve. Compared to an approach that applies lasso regularization on each of the SEMs, which is a special case of our method, the product regularization approach achieves lower mean squared error in estimating the mediation effects.

email: *yi_zhao@brown.edu*

40f. INFERENCE OF GENETIC NETWORK FROM NEXT GENERATION SEQUENCING DATA

Bochao Jia*, University of Florida

Faming Liang, University of Florida

Gaussian graphical models are widely used in determining genes network and their interactions. However, they are only optimal for modeling networks based on the data following the normal distribution. As there are varieties of high-dimensional count data in gene expression data such as mRNA, MicroRNA, and Copy Number, we developed a new strategy to deal with the count data for estimating high-dimensional gene networks. Unlike many appeared methods only for local graphical models, our method is a global one which can be convincing. We first assume the count data following the mixed Poisson distribution and transform them into continuous data through bayesian method. Then we normalize the data and calculate an equivalent measure of their partial correlation coefficients for Gaussian Graphical Models and proof that it is more powerful for inferring the gene interaction structure both in theoretical and numerical ways.

email: *jbc409@ufl.edu*

40g. INTERPRETABLE HIGH-DIMENSIONAL INFERENCE VIA SCORE MAXIMIZATION WITH AN APPLICATION IN NEUROIMAGING

Simon N. Vandekar*, University of Pennsylvania

Russell T. Shinohara, University of Pennsylvania

In the fields of neuroimaging and genetics there is interest in testing the association of a categorical or continuous outcome with a high-dimensional imaging or genetic variable. Often times several summary measures of the high-dimensional variable are created to sequentially test and localize the association with the outcome. In some cases the results for the summary measures are significant, but subsequent tests used to localized differences are underpowered and do not identify regions associated with the outcome. We propose a modification of Rao's score test that maximizes the score statistic in a linear subspace of the parameter space. If the test rejects the null hypothesis, then inference can be performed on the scores in the high-dimensional space by projecting the scores to the subspace where the score test was performed. This allows for inference in the high-dimensional space, which can be used to localize the association with the outcome, to be performed on the same degrees of freedom as the score test of association. We illustrate the method using cortical thickness data from the Alzheimer's Disease Neuroimaging Initiative where the results offer improved interpretability over performing sequential tests. Simulation results demonstrate that the test has competitive power relative to others used in neuroimaging and genetics.

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

40h. SINGLE INDEX LATENT FACTOR 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

The aim of this paper is to develop a single-index latent factor modeling (SILFM) framework to build an accurate prediction model for clinical outcomes based on a massive number of features. We develop a three-stage estimation procedure to build the prediction model. SILFM uses an independent screening method to select a set of informative features, which may have a complex nonlinear relationship with the outcome variables. Moreover, we develop a latent factor model to project all informative predictors onto a small number of local subspaces, which lead to a few key features that capture reliable and informative covariate information. Finally, we fit kernel ridge regression to those key features in order to accurately predict clinical outcomes. We systematically investigate the theoretical properties of SILFM, such as risk bounds and selection consistency. Our simulation results and real data analysis show that SILFM outperforms many state-of-the-art methods in terms of prediction accuracy.

email: *hojiny0504@gmail.com*

40i. MIXED MODELS FOR ORDINAL OUTCOMES IN TWIN AND SIBLING STUDIES WITH HIGH-DIMENSIONAL COVARIATE SPACES

Amanda E. Gentry*, Virginia Commonwealth University

Kellie J. Archer, Virginia Commonwealth University

The ongoing Brisbane Longitudinal Twin Study (BLTS) is being conducted in Australia and is funded by the US National Institute on Drug Abuse (NIDA). Adolescent twins and their non-twin siblings were sampled as a part of this study. We are analyzing a subset of this data that includes demographics, cannabis use metrics, and imputed genotypes for 8,572,909 single nucleotide polymorphisms (SNPs) for 1,307 patients. Our primary goal is to determine what combination of demographic information and SNPs may predict cannabis use, measured on an ordinal scale as: tried, tried but did not use frequently, used frequently. To conduct this analysis, we have extended the ordinal Generalized Monotone Incremental Forward Stagewise (GMIFS) method for mixed models. Our mixed model includes a random intercept term for each family with correlations between family members assigned according to a kinship

matrix; variance of this random term estimates genomic component of variation. Two additional random terms for monozygotic (identical) and dizygotic twins/siblings are estimated and the variances of these terms identify environmental contributions to the variance. Since the number of covariates is much greater than the number of patients, we use a stepwise procedure to achieve a parsimonious model. Our model is penalized so that only those SNPs which contribute significantly to the outcome will be allowed to enter the model. Our model allows demographic variables to be forced into the model without penalization.

email: *gentryae@vcu.edu*

40j. EVOLVING BAYESIAN NETWORKS: APPLICATIONS TO GENOMIC PATHWAYS AND LEARNING MODULES

Riten Mitra*, University of Louisville

Yuan Ji, NorthShore University Health System

Peter Mueller, University of Texas, Austin

We propose a class of hierarchical Bayesian models that simultaneously infers subgroups and subgroup specific biological networks in a patient population. The current paradigm of personalized medicine mostly rests on testing marginal differential expressions of some selected biomarkers. We extend this frontier by including semi-supervised learning of pathway interactions, integrating multimodal omics data, and clustering patients based on discovered pathways. We illustrate the use of our methods in discovering unknown cancer subtypes. We further build upon this clustering approach to model non-exchangeable networks evolving across time. These models are potentially applicable to brain networks in subjects undergoing specific motor tasks. The networks typically represent functional connectivity between neural units, and their evolution informs us about learning mechanisms. A novel class of Bayesian priors is proposed to regulate the smoothness of these evolving networks.

email: *riten82@gmail.com*

40k. SPARSE GROUP LASSO AND SVM WITH OVERLAPPING GROUPS

David Degras*, DePaul University

The incorporation of overlapping groups in penalization methods such as lasso and SVM has received substantial interest in the recent years. Indeed, overlapping groups naturally arise in areas like bioinformatics, e.g., when genes are known to belong to multiple functional groups. In comparison to gene-level penalization only, combined gene- and group-level penalization in regression/classification tasks has shown the promise of significant increases in efficiency in translational research. However, the state-of-the art penalization methods can either handle overlapping groups or induce sparsity both at the covariate and group level, but not both. In this work we combine these two features in a unified framework and aim to derive efficient algorithms for computing the solutions. To this end we will seek to extend to extend existing methods such as sparse group lasso and variational methods.

email: *ddegrasv@depaul.edu*

41. HIGH-THROUGHPUT EXPRESSION LANDSCAPE: WHAT'S NEXT FOR METHODS?

OVERCOMING BIAS AND BATCH EFFECTS IN RNA-Seq DATA

Michael I. Love, Dana-Faber Cancer Institute and Harvard School of Public Health

Rafael A. Irizarry*, Dana-Faber Cancer Institute and Harvard School of Public Health

In this talk I will demonstrate the presence of bias, systematic error and unwanted variability in RNA-sequencing data. I will show the substantial downstream effects these have on downstream results and how they can lead to misleading biological conclusions. Specifically, I will show how sequence bias can lead to incorrect alternative splicing estimates. I will do this using data from the public repositories as well as our own. We will then describe a solutions to these problems.

email: *rafa@jimmy.harvard.edu*

INTEGRATIVE MODELS FOR PREDICTING THE REGULATORY IMPACT OF RARE NON-CODING VARIATION

Alexis Battle*, Johns Hopkins University

The increase in availability of whole genome sequences (WGS) presents opportunity for understanding the impact of rare genetic variants. However, we are still limited in our ability to predict consequences of rare variants in non-coding regions of the genome. Diverse genomic annotations such as epigenetic data have been shown to be informative regarding regulatory elements, but are only moderately predictive of impact for individual variants. RNA-seq and other molecular data can provide evidence of cellular disruption on an individual basis, complementing genome sequencing. We have developed a Bayesian machine learning approach that integrates WGS with RNA-seq from the same individual along with diverse genomic annotations, performing joint inference to identify likely rare regulatory variants. We have applied this model to hundreds of WGS and corresponding RNA-seq samples and prioritized likely regulatory variants for each individual. We demonstrate that integrative models perform significantly better than predictions from WGS or RNA-seq

alone. Our probabilistic model offers great potential for identifying functional non-coding variants from individual genomes.

email: *ajbattle@cs.jhu.edu*

ANNOTATION-AGNOSTIC DIFFERENTIAL EXPRESSION ANALYSIS

Leonardo Collado-Torrres*, Johns Hopkins University

Alyssa Frazee, Johns Hopkins University

Michael I. Love, Dana-Farber Cancer Institute and Harvard School of Public Health

Rafael A. Irizarry, Dana-Farber Cancer Institute and Harvard School of Public Health

Andrew Jaffe, Lieber Institute for Brain Development

Jeffrey Leek, Johns Hopkins University

Differential expression (DE) analysis of RNA sequencing data typically relies on reconstructing transcripts or counting reads that overlap features, which depends on the known transcriptome. We previously introduced an intermediate statistical approach called differentially expressed region (DER) finder [1] that seeks to identify regions of the genome showing DE signal at base resolution. Using this approach we identified DERs associated with development and aging of the human brain, highlighting its incomplete annotation [2]. Here we describe software built around the DER approach to RNA-seq analysis. We introduce derfinder2 that allows for: (1) a computationally efficient annotation-agnostic DER finder called expressed-region analysis, (2) genome-scale analyses in a large number of samples, (3) flexible statistical modeling, including multi-group and time course analyses, and (4) data visualization in R. Our software permits a comprehensive analysis of RNA-seq data at base resolution, from preprocessing, to modeling, to annotation and visualization. We perform a complete comparison to feature counts based methods and demonstrate that expressed-region analysis sacrifices a small amount of power to enable discovery. Finally we apply this approach to public RNA-seq data from the developing human brain. The software is available from github.com/leekgroup/derfinder2. [1] Frazee et al, doi:10.1093/biostatistics/kxt053 [2] Jaffe et al, doi:10.1038/nn.3898.

email: *lcollado@jhu.edu*

DETECTING DIFFERENTIAL USAGE OF EXONS USING RNA-Seq DATA

Alejandro Reyes*, European Molecular Biology Laboratory

Simon Anders, Institute for Molecular Medicine Finland

Wolfgang Huber, European Molecular Biology Laboratory

The understanding of transcript isoforms and their functional differences often relies in transcriptome comparisons between biological contexts, such as different tissues or upon perturbation experiments. High-throughput sequencing of RNA (RNA-seq) provides the means to study alternative transcript isoform regulation. I will present DEXSeq, a method to test for differences in exon usage using RNA-seq. DEXSeq starts by summarizing experiments in count matrices. It models such counts using a Negative Binomial distribution. Generalized linear models are used for testing, which allow incorporating information from additional technical (e.g., batch information) or biological variables. The method provides reliable control of type I error by taking into account biological variation. The method also provides functions to visualize the results of the test. DEXSeq is implemented in an R/Bioconductor package.

email: *alejandro.reyes@embl.de*

42. STATISTICAL ISSUES IN ESTIMATING HEALTH DISPARITIES USING COMPLEX SAMPLES

COMPARING METHODS OF HEALTHCARE DISPARITY ESTIMATION IN THE PRESENCE OF COMPLEX SURVEY DESIGN

Benjamin Cook*, Harvard Medical School

Alan Zaslavsky, Harvard Medical School

To implement the Institute of Medicine definition of healthcare disparities, it is necessary to statistically adjust for racial/ethnic differences due to clinical appropriateness, need and patient preferences, but not differences due to discrimination or differential treatment due to an individual's insurance or socioeconomic status. Methods have been developed to implement this definition that use rank and propensity score based adjustment schemes to adjust for some variables but not others. The unbiasedness and consistency of variance estimators for these methods in the context of complex survey designs has yet to be evaluated. Data are from the 2004-2013 Medical Expenditure Panel Survey. We compare estimates and standard errors using a rank and propensity score method of implementing the IOM definition of disparity. To generate standard errors, we apply bootstrap (with and without accounting for the complex survey design) and balanced repeated replication (BRR) methods. We apply these methods to two outcomes, the use of beta blockers after AMI and the receipt of mental health care, the latter we expect to be more heavily influenced by clustering given its greater geographic variation. Results will provide researchers more information about optimal methods for measuring health care disparities in the presence of complex survey sampling.

email: *bcook@cha.harvard.edu*

EXTENSION OF THE PETERS-BELSON METHOD TO ESTIMATE HEALTH DISPARITIES AMONG MULTIPLE GROUPS USING LOGISITIC REGRESSION WITH SURVEY

Yan Li*, University of Maryland, College Park

Barry I. Graubard, National Cancer Institute

Pengyu Huang, University of Maryland, College Park

Joe Gastwirth, George Washington University

Determining the extent of a disparity, if any, between groups of people, for example, race or gender, is of interest in many fields, including public health for medical treatment and prevention of disease. An observed difference in the mean outcome between an advantaged group (AG) and disadvantaged group (DG) can be due to differences in the distribution of relevant covariates. The Peters-Belson (PB) method fits a regression model with covariates to the AG to predict, for each DG member, their outcome measure as if they had been from the AG. The difference between the mean predicted and the mean observed outcomes of DG members is the (unexplained) disparity of interest. We focus on applying the PB method to estimate the disparity based on binary/multinomial/proportional odds logistic regression models using data collected from complex surveys with more than one DG. Estimators of the unexplained disparity, an analytic variance-covariance estimator that is based on the Taylor linearization variance-covariance estimation method, as well as a Wald test for testing a joint null hypothesis of zero for unexplained disparities between two or more minority groups and a majority group, are provided. Simulation studies with data selected from simple random sampling and cluster sampling, as well as the analyses of disparity in body mass index in the National Health and Nutrition Examination Survey 1999-2004, are conducted. Empirical results indicate that the Taylor linearization variance-covariance estimation is accurate and that the proposed Wald test maintains the nominal level.

Email: *yli6@umd.edu*

EXAMINING SOCIOECONOMIC HEALTH DISPARITIES USING A RANK-DEPENDENT RÉNYI INDEX

Makram Talih*, Centers for Disease Control and Prevention

The Rényi index (RI) is a one-parameter class of indices that summarize health disparities among population groups by measuring divergence between the distributions of disease burden and population shares of these groups. The rank-dependent RI is a two-parameter class of health disparity indices that also accounts for the association between socioeconomic rank and health; it may be derived from a rank-dependent social welfare function. Two competing classes are discussed and the rank-dependent RI is shown to be more robust to changes in the distribution of either socioeconomic rank or health. The standard error and sampling distribution of the rank-dependent RI are evaluated using linearization and resampling techniques, and the methodology is illustrated using health survey data from NHANES and registry data from SEER. Such data underlie many population-based objectives within Healthy People 2020. The rank dependent RI provides a unified mathematical framework for eliciting various societal positions with regards to the policies that are tied to such wide-reaching public health initiatives. For example, if population groups with lower socioeconomic position were ascertained to be more likely to utilize costly public programs, then the parameters of the RI could be selected to reflect prioritizing those population groups for intervention or treatment.

email: *mtalih@cdc.gov*

ESTIMATING THE RELATIVE CONCENTRATION INDEX FROM COMPLEX SURVEY SAMPLES

Mandi Yu*, National Cancer Institute, National Institutes of Health

Benmei Liu, National Cancer Institute, National Institutes of Health

Yan Li, University of Maryland, College Park

The relative concentration index (RCI) is a widely used and attractive measure for measuring socioeconomic inequalities in health. First proposed and used by economist Nanak Kakwani in 1977, this index has its sampling distribution derived in 1997 also by Kakwani based on the assumption that the data comes from simple random samples. Recently, health surveys are increasingly recognized as valuable sources for assessing disparity for a wide range of health behaviors. However, estimators of RCI that incorporate complex samplings features, such as stratification, clustering, or unequal probability sampling, are not available for use in making valid inference about the direction and magnitude of relative disparity across socioeconomic groups. In this presentation, we derived and evaluated point and variance estimators of RCI under various complex sampling designs. We used a linearization approximation approach of deriving its variance estimators. We finally demonstrated its use in examining the disparity of breast cancer screening rate by a few socioeconomic indicators using health survey data from the Health Related Behaviors Survey of Active Duty Military Personnel and the National Health Interview Survey.

email: *yum3@mail.nih.gov*

43. STATISTICAL METHODS FOR NEUROSCIENCE

MULTI-SCALE FACTOR ANALYSIS OF HIGH DIMENSIONAL TIME SERIES DATA WITH APPLICATIONS TO fMRI

Hernando Ombao*, University of California, Irvine

Yuxiao Wang, University of California, Irvine

Chee-Ming Ting, Universiti Teknologi Malaysia

The difficulty in modeling functional connectivity is primarily due to the sheer high dimensionality of fMRI data. To address this problem, we develop a multi-scale factor analysis (MSFA) model which is a statistically principled approach to modeling and estimating resting state connectivity. The first step in our proposed framework is to reduce dimensionality by applying principal components analysis (PCA) within each anatomically parcellated region of interest (ROI). This dimension reduction approach is ideal for modeling connectivity because it summarizes localized activity by selecting components series that best explain localized (within-ROI) variance. The second step is to model connectivity between the ROIs or system networks by computing the dependence measure between components series extracted from each of the ROIs. In this paper, we measure connectivity by the RV-coefficient and other spectral-directed measures which is derived from the covariance and cross-spectrum between the components from the ROIs or networks. The dimension reduction step in our method differs from the most common approach which simply extracts the average time series in a ROI. This approach is not optimal because a single summary time series is not likely to sufficiently capture localized brain activity. The proposed procedure simultaneously accomplishes the following desired goals: (1.) it gives a representation of localized brain activity that is an optimal solution to the PCA criterion of maximizing the explained variation within each ROI; (2.) it captures the multi-scale dependence structure at both local (within-ROI) level and global (between ROIs and between networks) level; and (3.) it achieves dimension reduction therefore can efficiently handle the massive fMRI data. The novel MSFA approach is used to study functional connectivity in resting-state fMRI data, which reveals interesting modular and hierarchical structure of human brain networks.

email: *hombao@uci.edu*

KINEMATIC DATA IN MOTOR CONTROL EXPERIMENTS

Jeff Goldsmith*, Columbia University

Tomoko Kitago, Columbia University Medical Center

Stroke is the leading cause of long-term disability in the United States, with an incidence approaching one million events each year. Experiments involving kinematic data -- dense recordings of hand or finger position over time during the execution of a motion -- can provide deep insights into the neurological processes underlying disability induced by stroke. We take a functional data approach to the analysis of kinematics by posing a bivariate function-on-scalar regression with subject-level random functional effects. We express fixed effects and random effects using penalized splines; parameters are jointly estimated in a Bayesian framework using both MCMC and a computationally efficient variational approximation. Application results indicate that the effect of stroke on motor control has a systematic component observed across subjects.

email: *jeff.goldsmith@columbia.edu*

MULTIVARIATE PATTERN ANALYSIS AND CONFOUNDING IN NEUROIMAGING

Kristin Linn*, University of Pennsylvania

Bilwaj Gaonkar, University of Pennsylvania

Jimit Doshi, University of Pennsylvania

Christos Davatzikos, University of Pennsylvania

Russell Shinohara, University of Pennsylvania

Neuroimaging studies often quantify disease-related structural brain differences between populations using a multivariate pattern analysis (MVPA) such as the support vector machine (SVM). The SVM is trained to discriminate between groups, and the weights indicate which brain regions jointly drive the discriminative rule. However, classifier training in the presence of confounders may lead to identification of false disease patterns and spurious results. This occurs when classifiers rely heavily on regions that are strongly correlated with the confounders instead of regions that encode subtle disease changes. The imaging literature recommends using parametric models to regress out confounder effects at each brain region before SVM training. We show that this approach does not properly address the issue of confounding in MVPA. Instead, we propose a novel method that incorporates inverse probability weighting (IPW) during classifier training.

email: *klinn@upenn.edu*

A BAYESIAN APPROACH TO THE STUDY OF DYNAMIC FUNCTIONAL CONNECTIVITY NETWORKS IN fMRI DATA

Michele Guindani*, University of Texas MD Anderson Cancer Center

Ryan Warnick, Rice University

Marina Vannucci, Rice University

Erik Erhardt, University of New Mexico

Elena Allen, MRN Mind Research Network and University of New Mexico

Vince Calhoun, MRN Mind Research Network and University of New Mexico

fMRI studies have traditionally assumed stationarity of the connectivity patterns observed in a subject during a fMRI experiment. While the assumption has successfully allowed to study large-scale properties of brain functioning, it is generally recognized that functional connectivity varies with time and tasks performed. We describe a novel Bayesian methodological framework for the analysis of temporal dynamics of functional networks in task-based fRMi data collected on a single subject. Our proposed formulation allows joint modeling of the task-related activations in addition to the dynamics of individual functional connectivity. Furthermore, we allow simultaneous learning of the common and differential edges (interactions) in the inferred time-varying functional networks. We illustrate the proposed approach by means of simulation and an analysis on a real fMRI dataset.

email: *mguindani@mdanderson.org*

44. RECENT ADVANCES IN STATISTICAL METHODS FOR GENETIC EPIDEMIOLOGY

RARE VARIANT ASSOCIATION TESTS WITH LONGITUDINAL OUTCOME DATA

Zihuai He, University of Michigan

Seungeung Lee, University of Michigan

Min Zhang, University of Michigan

Bhramar Mukherjee*, University of Michigan

In this talk we propose a generalized score test with small sample correction for rare variants association analysis with longitudinal outcome. We compare the test with burden/collapsed tests in the generalized estimating equations framework as well as sequential kernal association tests (SKAT and SKAT-O) using average outcome. We also propose a weighted combination of our proposed test and collapsed/burden test similar in spirit to SKAT-O. The methods are illustrated by using Exome-chip data from the Multi-Ethnic Study of Atherosclerosis.

email: *bhramar@umich.edu*

DETECTING ASSOCIATIONS OF RARE VARIANTS WITH COMMON DISEASES USING SNP DATA ON FAMILIES

Shili Lin*, The Ohio State University

Meng Wang, Nationwide Children's Hospital

In recent years, a myriad of new statistical methods have been proposed for detecting associations of rare single-nucleotide variants (SNVs) with common diseases. These methods can be classified as "collapsing", or "haplotyping" based, with the former composed of most of the methods proposed to date. However, recent works have suggested that haplotyping-based methods may offer advantages and can even be more powerful than collapsing methods in certain situations. We propose a family-based logistic Bayesian Lasso (famLBL) method that is designed to estimate effects of haplotypes using common SNV data on families. By choosing appropriate prior distributions, effect sizes of unassociated haplotypes can be shrunk toward zero, allowing for more precise estimation of rare and common associated haplotypes, thereby achieving greater detection power. We evaluate famLBL and compare its performance with another haplotyping (but population-based) method and three collapsing methods, both population-based and family-based. The results show that haplotyping methods can be more powerful than collapsing methods if there are interacting SNVs leading to larger haplotype effects. Even if only common SNVs are genotyped, haplotype methods can still detect specific rare haplotypes that tag rare causal SNVs. As expected, family-based methods are robust, whereas population-based methods are susceptible, to population substructure.

email: *shili@stat.osu.edu*

DETECTION OF SET-BASED GENE-ENVIRONMENT INTERACTIONS IN FAMILIES

Saonli Basu*, University of Minnesota

Brandon Coombes, University of Minnesota

The development of a complex trait is an intricate dynamic process controlled by a network of genes and environmental factors. In recent years, the availability of high throughput genomic data has generated ample interests in investigating the complex interplay between these genes and environmental factors (G-E interaction). One way to increase power for detection of G-E interaction is to improve the effect size(s) by aggregating the single-nucleotide polymorphisms (SNPs) within a gene in what we call SNP-sets. We propose here a test for detection of interaction between a SNP-set and a group of correlated environmental factors in families by using a likelihood-based dimension reduction approach within a random-effect model framework. The proposed approach employs a scoring system to capture the effect of a group of interacting SNPs and environmental exposures. We also extend several variance component based tests to study G-E interaction in families. We illustrate our model through simulation studies and compare the performance of different methods to detect G-E interaction. We demonstrate that the performance of these methods vary widely based on the directionality and sparsity of the interaction effects and our dimension reduction approach performs particularly well in presence of interaction effects in the same direction.

e-mail: *saonli@umn.edu*

ADDITIVE MODELS FOR EVALUATING PREDICTIVE BIOMARKERS IN CANCER EPIDEMIOLOGY STUDIES

Jaya M. Satagopan*, Memorial Sloan Kettering Cancer Center

There is considerable interest in finding predictive biomarkers that can guide treatment options for mutation carriers and non-carriers. This has led to an increased interest in the evaluation of gene-treatment and gene-gene interactions in epidemiology studies. Certain interactions arise due to the scale on which the outcome is measured. For binary outcomes, the scale refers to a link function. By choosing an appropriate link function, we can fit an additive model when the data satisfy certain properties. Equivalently, we can represent the interaction terms in a parsimonious manner when we model the data under a different link function. This talk will discuss methods for exploiting model parsimony under a clinically useful link function to evaluate predictive biomarkers in an efficient manner, and illustrate them using data from published cancer epidemiology studies.

email: *satagopj@mskcc.org*

45. RECENT ADVANCES IN SURVIVAL ANALYSIS WITH HIGH-DIMENSIONAL DATA

FEATURE SCREENING IN ULTRAHIGH DIMENSIONAL COX'S MODEL

Guangren Yang, Jinan University

Ye Yu, Wells Fargo Bank

Runze Li*, The Pennsylvania State University

Anne Buu, University of Michigan

Survival data with ultrahigh dimensional covariates such as genetic markers have been collected in medical studies and other fields. In this work, we propose a feature screening procedure for the Cox model with ultrahigh dimensional covariates. The proposed procedure is distinguished from the existing sure independence screening (SIS) procedures (Fan, Feng and Wu, 2010, Zhao and Li, 2012) in that the proposed procedure is based on joint likelihood of potential active predictors, and therefore is not a marginal screening procedure. The proposed procedure can effectively identify active predictors that are jointly dependent but marginally independent of the response without performing an iterative procedure. We develop a computationally effective algorithm to carry out the proposed procedure and establish the ascent property of the proposed algorithm. We further prove that the proposed procedure possesses the sure screening property. That is, with the probability tending to one, the selected variable set includes the actual active predictors. We conduct Monte Carlo simulation to evaluate the finite sample performance of the proposed procedure and further compare the proposed procedure and existing SIS procedures. The proposed methodology is also demonstrated through an empirical analysis of a real data example.

email: *rzli@psu.edu*

INTEGRATING MULTIDIMENSIONAL OMICS DATA FOR CANCER PROGNOSIS

Shuangge Ma*, Yale University

Prognosis is of essential interest in cancer research. Multiple types of omics measurements "including mRNA gene expression, methylation, copy number variation, SNP, and others" have been implicated in cancer prognosis. The analysis of multidimensional omics data is challenging because of the high data dimensionality and, more importantly, because of the interconnections between different units of the same type of measurement and between different types of omics measurements. In our study, we have developed novel regularization-based methods, effectively integrated multidimensional data, and constructed prognosis models. It is shown that integrating multidimensional data can lead to biological discoveries missed by the analysis of one-dimensional data and superior prognosis models.

email: *shuangge.ma@yale.edu*

SURVIVAL PREDICTION FROM LARGE-SCALE DATA USING METRIC LEARNING

Daniel Conn, University of California, Los Angeles

Christina Ramirez, University of California, Los Angeles

Zhenqiu Liu, Cedars-Sinai Medical Center

Gang Li*, University of California, Los Angeles

We consider the problem of building a survival prediction model based on large-scale data. Standard regression models such as the Cox model are often inadequate to describe complex relations and interactions that may be present in a large heterogeneous population. This paper introduces a new approach to building a survival prediction model by adapting the metric learning methodology to a censored outcome. The method is an extension of kernel regression designed to overcome the flaws of standard nonparametric regression methods in higher dimensions. It uses data to learn a kernel function that adaptively down-weights unimportant features, up-weights important features, achieves dimension reduction in a supervised way. It effectively handles nonlinear relations, complex interactions, and highly correlated features. We demonstrate the usefulness of our method in data rich settings using both simulated and real data.

email: *vli@ucla.edu*

46. DISSECTING MULTIPLE IMPUTATION FROM A MULTI-PHASE INFERENCE PERSPECTIVE

DISSECTING MULTIPLE IMPUTATION FROM A MULTI-PHASE INFERENCE PERSPECTIVE: WHAT HAPPENS WHEN GOD'S, IMPUTER'S AND ANALYST'S MODELS ARE UNCONGENIAL?

Xianchao Xie, Two Sigma

Xiaoli Meng*, Harvard University

Real-life data are almost never really real. By the time the data arrive at an investigator's desk or disk, the raw data have likely gone through some cleaning processes, such as standardization, recalibration, and imputation. Dealing with such a reality scientifically requires a more holistic multi-phase perspective. This article provides an in-depth look from this broader perspective, into multiple-imputation (MI) inference (Rubin (1987)) under uncongeniality (Meng (1994)). We present a general estimating-equation decomposition theorem, resulting in an analytic (asymptotic) description of MI inference as an integration of the knowledge of the imputer and the analyst. These results help to reveal how the quality of and relationship between the imputer's model and analyst's procedure affect MI inference, including how a seemingly perfect procedure under the "God-versus-me" paradigm is actually inadmissible when God's, imputer's, and analyst's models are uncongenial to each other. Our theoretical investigation also leads to procedures that are as trivially implementable as Rubin's combining rules, yet with guaranteed confidence coverages under any degree of uncongeniality.

email: *meng@stat.harvard.edu*

DISCUSSANTS:

Trivellore E. Raghunathan, University of Michigan

Jerry Reiter, Duke University

Tony Desmond, University of Guelph

47. INNOVATIVE CLINICAL TRIAL DESIGN AND ANALYSIS METHODS

BAYESIAN DESIGN OF SUPERIORITY CLINICAL TRIALS FOR RECURRENT EVENTS DATA WITH APPLICATIONS TO BLEEDING AND TRANSFUSION EVENTS IN MYELODYPLASTIC SYNDROME

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

Ming-Hui Chen, University of Connecticut

Donglin Zeng, University of North Carolina, Chapel Hill

Kuolung Hu, Amgen, Inc.

Catherine Jia, Amgen, Inc.

In many biomedical studies, patients may experience the same type of recurrent event repeatedly over time, such as bleeding, multiple infections and disease. In this paper, we propose a Bayesian design to a pivotal clinical trial in which lower risk myelodysplastic syndromes (MDS) patients are treated with MDS disease modifying therapies. One of the key study objectives is to demonstrate the investigational product (treatment) effect on reduction of platelet transfusion and bleeding events while receiving MDS therapies. In this context, we propose a new Bayesian approach for the design of superiority clinical trials using recurrent events frailty regression models. Historical recurrent events data from an already completed phase 2 trial are incorporated into the Bayesian design via the partial borrowing power prior of Ibrahim et al. (2012). An efficient Gibbs sampling algorithm, a predictive data generation algorithm, and a simulation-based algorithm are developed for sampling from the fitting posterior distribution, generating the predictive recurrent events data, and computing various design quantities such as the type I error rate and power, respectively. An extensive simulation study is conducted to compare the proposed method to the existing frequentist methods and to investigate various operating characteristics of the proposed design.

email: *ibrahim@bios.unc.edu*

STATISTICAL METHODS FOR CONDITIONAL SURVIVAL ANALYSIS

Sin-Ho Jung*, Duke University

Sunkyu Choi, Samsung Medical Center

Ho Yun Lee, Samsung Medical Center

We investigate the survival distribution of patients who have survived over a certain time period. This is called a conditional survival distribution. In this talk, we show that one-sample estimation, two-sample comparison and regression analysis of conditional survival distributions can be conducted using the regular methods for unconditional survival distributions that are provided by the standard statistical software, such as SAS and SPSS. We will prsent results from extensive simulations to evaluate the finite sample property of these conditional survival analysis methods. We will also illustrate these methods with real clinical data.

email: *sinho.jung@duke.edu*

MULTI-ARM PLATFORM DESIGNS FOR SCREENING EFFECTIVE TREATMENTS VIA PREDICTIVE PROBABILITY

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

Brian P. Hobbs, University of Texas MD Anderson Cancer Center

Nan Chen, University of Texas MD Anderson Cancer Center

The process of screening effective treatments one-at-a-time is inefficient and costly. We introduce a statistical framework for designing

and conducting randomized multi-arm screening trials using the concept of platform designs and Bayesian predictive probability. In essence, the proposed platform-based approach consolidates inter-study control arms enabling investigators to assign more new patients to novel therapies. The process accommodates mid-trial modifications to the study arms that allow both dropping poorly performing agents as well as incorporating new agents. Compared to randomized two-arm trials, screening platforms have the potential to yield considerable reductions in cost, alleviate the bottleneck between phase I and II, eliminate bias stemming from inter-trial heterogeneity, and control for multiplicity over a sequence of a priori planned studies, yet, without sacrificing frequentist properties for comparing treatments. The gains in efficiency facilitated by platform-based designs could be substantial in oncologic settings, wherein trials often suffer from low enrollment, an unacceptably high rate of failure in phase III, and long inter-trial latency periods. Simulations are provided to compare the operating characteristics of the proposed multi-arm platform designs and the sequentially conducted trials.

email: *jjlee@mdanderson.org*

OPTIMAL FLEXIBLE SAMPLE SIZE DESIGN WITH ROBUST POWER

Lu Cui*, AbbVie Inc.

Lanju Zhang, AbbVie Inc.

Bo Yang, AbbVie Inc.

It is well recognized that sample size determination is challenging due to the uncertainty on the treatment effect size. Flexible sample designs, including popular group sequential, promising-zone, and sample size re-estimation designs, are proposed as alternatives to fixed sample size design. Different opinions favoring one type over the other exist. We propose an approach using an appropriate optimality criterion to select the best design among all the candidate designs. Our results show that (1) for the same type of design, for example group sequential designs, there is a room for significant improvement through our optimization approach ; (2) Optimal promising zone design appears having no advantages over optimal group sequential design; and (3) Optimal design with sample size re-estimation delivers the best adaptive performance. We conclude that to deal with the challenge of sample size determination due to uncertainty of the projected treatment effect, an optimization approach can help to select the best flexible sample design that provides most robust power over a range of treatment effect size with an efficient average sample size.

email: *lu.cui@abbvie.com*

48. STATISTICAL ADVANCES IN EVOLUTIONARY DYNAMICS OF INFECTIOUS DISEASES

ALGORITHMS LINKING PHYLOGENETIC AND TRANSMISSION TREES FOR MOLECULAR INFECTIOUS DISEASE EPIDEMIOLOGY

Eben Kenah*, University of Florida

Tom Britton, Stockholm University

M. Elizabeth Halloran, Fred Hutchinson Cancer Research Center and University of Washington

Ira M. Longini, Jr., University of Florida

Recent work has considered the use of densely-sampled genetic data to reconstruct the transmission trees in outbreaks. Because transmission trees from one outbreak do not generalize to future outbreaks, scientific insights useful for public health are more likely to be obtained by using genetic data to estimate transmission parameters more precisely. In a survival analysis framework, parameter estimation is based on sums or averages over possible transmission trees. By restricting the set of possible trees, a phylogeny can increase the efficiency of these estimates. The leaves of the phylogeny represent sampled pathogens, which have known hosts. The interior nodes represent common ancestors of sampled pathogens, which have unknown hosts. We show that there is a one-to-one relationship between the possible assignments of interior node hosts and the transmission trees consistent with the phylogeny and the epidemiologic data. We develop algorithms to find the set of possible hosts at each interior node and to generate all possible transmission trees given these host sets. We apply these methods to foot-and-mouth disease virus outbreaks in the United Kingdom in 2001 and 2007.

email: *ekenah@ufl.edu*

PHYLODYNAMIC ANALYSIS WITH LIMITED DATA: EMERGENCE AND EPIDEMIOLOGICAL IMPACT OF TRANSMISSIBLE DEFECTIVE DENGUE VIRUSES

Ruian Ke, North Carolina State University

John Aaskov, Queensland University of Technology

Edward C. Holmes, University of Sydney

James O. Lloyd-Smith*, University of California, Los Angeles

Intra-host sequence data from RNA viruses have revealed the ubiquity of defective viruses in natural viral populations. The discovery of a transmissible lineage of defective dengue virus type 1 (DENV-1) in Myanmar, first seen in 2001, raised questions about

transmissible defective viruses and their epidemiological impact. By combining phylogenetic analyses and dynamical modeling, we investigate how processes at intra-host and inter-host scales shaped the emergence and spread of the defective DENV-1 lineage. We show that the defective virus was transmitted primarily through co-transmission with the functional virus, and that, surprisingly, this co-transmission is more efficient than transmission of functional dengue viruses alone. This implies that the defective lineage should increase overall incidence of dengue infection, which could explain the historically high dengue incidence in Myanmar in 2001-2002. Our results show the potential for defective viruses to impact the epidemiology of human pathogens, or to emerge as circulating infections in their own right. They also show that interactions between viral variants, such as complementation, can open new routes to viral emergence.

email: *jlloydsmith@ucla.edu*

AN EFFICIENT BAYESIAN INFERENCE FRAMEWORK FOR COALESCENT-BASED NONPARAMETRIC PHYLODYNAMICS

Shiwei Lan*, University of Warwick

Julia A. Palacios, Harvard University and Brown University

Michael Karcher, University of Washington

Vladimir N. Minin, University of Washington

Babak Shahbaba, University of California, Irvine

The field of phylodynamics focuses on the problem of reconstructing population size dynamics over time using current genetic samples taken from the population of interest. This technique has been extensively used in many areas of biology, but is particularly useful for studying the spread of quickly evolving infectious diseases agents, e.g.,\ influenza virus. Phylodynamic inference uses a coalescent model that defines a probability density for the genealogy of randomly sampled individuals from the population. When we assume that such a genealogy is known, the coalescent model, equipped with a Gaussian process prior on population size trajectory, allows for nonparametric Bayesian estimation of population size dynamics. While this approach is quite powerful, large data sets collected during infectious disease surveillance challenge the state-of-the-art of Bayesian phylodynamics and demand inferential methods with relatively low computational cost. To satisfy this demand, we provide a computationally efficient Bayesian inference framework based on Hamiltonian Monte Carlo for coalescent process models. Moreover, we show that by splitting the Hamiltonian function we can further improve the efficiency of this approach. Using several simulated and real datasets, we show that our method provides accurate estimates of population size dynamics and is substantially faster than alternative methods based on elliptical slice sampler and Metropolis-adjusted Langevin algorithm.

email: *S.Lan@warwick.ac.uk*

EFFECTS OF IGNORING RECOMBINATION IN PHYLODYNAMICS OF INFECTIOUS DISEASES

Julia A. Palacios*, Harvard University and Brown University

Phylodynamic analyses of infectious diseases are powerful tools that allow us to recover a relative measure of the number of infections through time known as effective population size-, from genomic samples. Phylodynamic analyses from genomic data at multiple loci usually assume that locus-specific genealogies are independent of each other. However, it is well known that recombination plays an important role in the observed genetic diversity, creating complex dependencies among genealogies at different loci. Here, we explore the effects of ignoring recombination when genealogies are partially linked on Bayesian nonparametric estimates of the effective population size trajectory. Through simulation of locus-specific genealogies under the Sequentially Markov Coalescent process (SMC), we assess the bias and the effect on accuracy of our estimates when ignoring recombination in our modeling framework. Our Bayesian nonparametric estimates of effective population size trajectories rely on Integrated Nested Laplace Approximations (INLA) previously developed for Phylodynamics. Finally, we provide a method for correction of bias and coverage of credible intervals.

email: *Julia.pal.r@gmail.com*

49. BAYESIAN SEMI-PARAMETRIC AND NON-PARAMETRIC METHODS

FLEXIBLE BAYESIAN SURVIVAL MODELING WITH SEMIPARAMETRIC TIME-DEPENDENT AND SHAPE-RESTRICTED COVARIATE EFFECTS

Thomas A. Murray*, University of Texas MD Anderson Cancer Center

Brian P. Hobbs, University of Texas MD Anderson Cancer Center

Daniel J. Sargent, Mayo Clinic

Bradley P. Carlin, University of Minnesota

Bayesian analysis of time-to-event data using a piecewise exponential model is common because it is reasonably flexible and easy to implement in popular Gibbs sampling software; however, it requires an unrealistic piecewise constant hazard assumption. We propose an easy to implement alternative that assumes a piecewise linear log-hazard. The proposed model uses a computationally convenient low-rank thin plate spline formulation for the log-hazard. We discuss extensions of the proposed model that facilitate estimating time-dependent and proportional-hazards (i.e., time-independent) covariate effects, possibly subject to shape restrictions. We show via simulation that our method can improve

estimation of important functions, e.g., log-hazards, survival distributions, and hazard ratios. We apply our method to analyze colorectal cancer data from a clinical trial comparing overall survival in two novel chemotherapy regimens relative to the standard of care. In this analysis, we estimate time-dependent hazard ratios comparing each novel regimen against the standard of care, while adjusting for a non-decreasing proportional-hazards effect of aspartate transaminase (a liver function biomarker).

email: *tamurray@mdanderson.org*

A BAYESIAN SEMIPARAMETRIC APPROACH FOR PANEL COUNT DATA

Jianhong Wang*, University of South Carolina, Columbia

Xiaoyan Lin, University of South Carolina, Columbia

In this paper, we propose a Bayesian semiparametric approach for analyzing panel count data under the proportional mean model. Specifically, Poisson process data are assumed and monotone I-splines are adopted to model the unknown baseline function. The regression parameters and the baseline function can then be simultaneously estimated. To facilitate the posterior computation, Poisson data augmentation is developed. The proposed Gibbs sampler is efficient and easy to implement because all of the full conditional distributions either have closed form or are log-concave. Extensive simulations are conducted to evaluate our proposed method, and the proposed approach has been compared with a parametric approach and an adapted Rosen approach through simulations. Our proposed method is also illustrated by a famous bladder tumor panel count data.

email: *w.jianhong@yahoo.com*

NONPARAMETRIC SMOOTHING ESTIMATION OF FECUNDABILITY FROM A CONCEPTION MODEL

Mohammed R. Chowdhury*, Kennesaw State University

Nonparametric estimation and inferences of fecundability have important applications in population studies and demographic research. We propose in this paper an estimation approach based on age-specific parametric models. We assume that the outcome variable at each given age follows a parametric model, but the parameters are smooth function of time (age). Our estimation is based on a two-step smoothing method, in which we first obtain the raw estimators of the parameters at a set of disjoint time points, and then compute the final estimators at any time by smoothing the raw estimators. Asymptotic properties, including the asymptotic biases, variances and mean squared errors, have been derived for the local polynomial smoothed estimators. Applications of our two-step estimation method have been demonstrated through a large demographic studies. Finite sample properties of our procedures are investigated through a simulation study.

email: *chowdhury@kennesaw.edu*

MARGINAL BAYESIAN HIERARCHICAL MODEL FOR MULTIVARIATE BINARY DATA TO ESTIMATE THE ETIOLOGY OF CHILDHOOD PNEUMONIA

Detian Deng*, Johns Hopkins University

Scott Zeger, Johns Hopkins University

Pneumonia, infection of the lung, is the number one cause of death for children under five. It is caused by more than 30 different pathogens. The PERCH Study is a multi-country case-control study to estimate the frequency with which each pathogen causes pneumonia (etiology distribution). This goal is challenging because sampling directly from a child's lung is not typically feasible. Rather pathogens are enumerated by PCR from multiple peripheral sites including the nose and blood. While previous methods only allow single-pathogen cause or fixed pairs of joint-cause, this talk will introduce a novel marginal Bayesian hierarchical model for multivariate binary data to estimate the etiology distribution, allowing for the possibility that multiple pathogens can cause a child's disease. This advantage is achieved by using an integrated likelihood that integrates the saturated model parameters over the feasible region set by marginal parameters of interests. It can be viewed as a bridge from the fully flexible log-linear model to the overly constrained multinomial model in which a single pathogen is assumed to be the cause. This model also features the novel partially informative prior that incorporates the knowledge about the likely number of pathogens that comprise sufficient cause.

email: *ddeng3@jhu.edu*

A SEMIPARAMETRIC BAYESIAN APPROACH TO BORROW INFORMATION FROM HISTORICAL CONTROL DATA IN TWO ARM CLINICAL TRIALS

Arpita Chatterjee*, Georgia Southern University

Historical information is always relevant for designing clinical trials. The incorporation of historical information in the new trial can be very benefitial. Some of these benefits include reduction of effective sample size, a significant increase in the statistical power, reduction of cost and ethical hazard. However, if current and historical trils conflict, borrowing information can give misleading results. In this project a semiparametric Bayesian method based on Dirichlet Process prior is introduced to borrow relevant information from historical control data. The scale parameter of the DP prior plays a crucial role by controlling the depndencies between the historical and current trials. The performances of the proposed

method is further compared with other competing methods in simulated data sets.

email: *achatterjee@georgiasouthern.edu*

BAYESIAN MULTIVARIATE NONLINEAR MIXED EFFECTS MODELS WITH A MATRIX STICK-BREAKING PROCESS PRIOR

Xiao Wu*, University of Florida

Michael J. Daniels, University of Texas, Austin

Analysis of longitudinal magnetic resonance imaging (MRI) and spectroscopy (MRS) data poses two challenges: accommodation of heterogeneity among multiple outcomes and clustering of similar (sub)units. For longitudinal MR muscle measures, heterogeneity in three aspects needs to be considered: among measures, among muscles, and across time; we want local subunit-specific clustering in which two patients may have the same effects for certain muscle/measure combinations but not others. Therefore, we propose a nonlinear mixed effects model to characterize fundamental components of disease progression and extend the matrix stick-breaking process by Dunson et al.(2008) for borrowing information across multidimensional outcomes and local clustering of similar subunits. The model is specified in a Bayesian framework and estimated through posterior computation using Markov chain Monte Carlo. The method is applied to a natural history study of Duchenne muscular dystrophy.

email: *xiaowu@ufl.edu*

BAYESIAN ADDITIVE PARTIAL LINEAR MODELS WITH MEASUREMENT ERROR AND HETEROSCEDASTIC REGRESSION ERROR VARIANCE

Chang Liu, University of Rochester

Sally W. Thurston*, University of Rochester

We consider the problem of estimation and inference for the additive partial linear model, when either a linear or a nonlinear covariate is measured with error. We address the situation in which the regression error variance is heteroscedastic and may also depend on the error-prone covariate. Penalized splines are used to estimate both the variance function and the unknown smooth functions in the outcome model, and the slice sampler is used to draw samples of the individual regression errors. Correction for measurement error bias and estimation of the smoothing parameters for the penalized splines are straightforward in our Bayesian hierarchical model. Simulation results show that explicitly modeling the variance function reduces the bias of the regression coefficient for the error prone covariate relative to the naive method.

email: *sally_thurston@urmc.rochester.edu*

50. BAYESIAN VARIABLE SELECTION

ALTERED SINGULAR BAYESIAN INFORMATION CRITERIA FOR BIVARIATE MIXTURE MODELS

Richard Charnigo*, University of Kentucky

Qian Fan, Wells Fargo

Ruriko Yoshida, University of Kentucky

Mathias Drton, University of Washington

Hongying Dai, Children's Mercy Hospital

The Bayesian Information Criterion (BIC) consistently estimates the order of a finite mixture model under mild conditions (Keribin, 2000). However, the BIC does not retain its interpretation as the approximate log marginal likelihood for irregular models (including finite mixtures), which led Drton and Plummer (2015) to introduce the singular Bayesian Information Criterion (sBIC). The SBIC is consistent and has the aforementioned interpretation under mild conditions. The present work has three aims. First, an appropriate penalty for selecting the order of an irregular model depends on the unknown truth, which appears to create a problem with circular reasoning. Drton and Plummer (2015) resolve this difficulty in one way, but we now propose three other means for addressing this problem; these give rise to three altered singular Bayesian Information Criteria (asBIC). Second, the sBIC has not yet been implemented for bivariate normal mixture models; we do so here, and we also establish the three aSBIC for these models. Finally, we apply the sBIC and the three asBIC to bivariate normal mixture models representing infant health data, to see whether and how conclusions differ from those that would be obtained via the BIC or the Akaike Information Criterion (AIC).

email: *RJCharn2@aol.com*

BAYESIAN BI-LEVEL VARIABLE SELECTION

Eunjee Lee*, University of North Carolina, Chapel Hill

Hongtu Zhu, University of North Carolina, Chapel Hill

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

A genome-wide association study (GWAS) focuses on identifying important SNPs to relate to clinical outcomes. Simple (and popular) GWASs conduct a number of marginal tests: examination of the effect of each SNP one by one. But they face two main challenges: dealing with multiple testing and accounting for dependency structure among SNPs. In order to resolve those limitations, we propose a Bayesian bi-level variable selection (BBVS) method in the accelerated failure time (AFT) model. It aims to detect SNPs associated with time to event outcomes by considering all the SNPs simultaneously

and incorporating their grouping information. Our method has two hierarchical levels of variable selection: the first one is group-wise and the second level is element-wise variable selection. First, we identify important groups of variables and update the censored event time by data augmentation. In the second level, we include variables of the selected groups as covariates in the AFT. To conduct element-wise variable selection, shrinkage priors are employed. In particular, we extend Dirichlet-Laplace shrinkage priors proposed by Bhattacharya et al. (2014) to incorporate the grouping information. We applied our proposed method to detect important genes and SNPs related to time to conversion from mild cognitive impairment to Alzheimer's disease.

email: *eunjee2@email.unc.edu*

BAYESIAN VARIABLE SELECTION FOR SKEWED HETEROSCEDASTIC RESPONSE

Libo Wang*, Florida State University

Yuanyuan Tang, AbbVie

Debajyoti Sinha, Florida State University

Debdeep Pati, Florida State University

Stuart Lipsitz, Brigham and Women's Hospital

In this article, we propose new Bayesian methods for selecting and estimating a sparse coefficient vector for skewed heteroscedastic response which are commonplace in medical cost studies. Our novel Bayesian procedures can handle response with large outliers, focus on evaluating median and other quantile functions, and asymptotically select the true set of predictors when the number of covariates increases in the same order of the sample size. Via a simulation study and a re-analysis of a medical cost study with large number of potential predictors, we illustrate the ease of implementation and other practical advantages of our approach compared to existing methods for such studies.

email: *l.wang@stat.fsu.edu*

BAYESIAN RANKING AND SELECTION WITH APPLICATION TO IDENTIFICATION OF RISK GENES

Xiaoqian Sun*, Clemson University

Feng Luo, Clemson University

Anand K. Srivastava, Greenwood Genetic Center

Identifying risk genes plays an important role in finding generic causes of autism spectrum disorder. We consider the integrated model of de novo mutations and transmitted variation, which depends on the mutation rate of gene, the population frequency of disease genotype variants, and the relative risk of mutation or variants. The risk genes correspond to those with large relative risks of mutation/variants and thus prioritizing or ranking genes based on their relative risks of mutation is an important statistical task. In this talk, we propose several Bayesian ranking and selection methods based on different ranking criteria, which all incorporate borrowing information across different genes in the study. The comparison of our ranking methods with some conventional methods is carried out through extensive simulation studies. Our proposed approaches are then applied to a dataset that combine case-control and family based studies in autism spectrum disorder.

email: *xsun@clemson.edu*

BAYESIAN VARIABLE SELECTION IN ADDITIVE PARTIAL LINEAR MODELS WITH ERROR IN VARIABLES

Chang Liu*, University of Rochester

Hongqi Xue, University of Rochester

Sally W. Thurston, University of Rochester

Adaptive LASSO and group LASSO are widely applied for variable selection in many types of statistical models. However, few papers focus on variable selection in additive partial linear models with measurement error. Within a Bayesian framework, we develop adaptive LASSO and group LASSO procedures for additive partial linear models when either a linear or nonlinear covariate is measured with error. It is well known that full Bayesian posterior inference can be expensive in computational efficiency for high-dimensional data and lacks an internal mechanism of dimension reduction for variable selection problems. Thus we propose a multiple-stage procedure to increase the efficiency of a full Bayesian posterior inference procedure. When applied to data with up to 100 linear and 20 nonlinear covariates, our simulation study shows that at least 95 of the linear covariates and 17 to 19 of the nonlinear covariates are identified correctly. We also find a substantial gain in efficiency of the multiple-stage procedure over the single-stage procedure.

email: *salvatore.cliu@gmail.com*

BAYESIAN FEATURE SCREENING FOR BIG NEUROIMAGING DATA VIA MASSIVELY PARALLEL COMPUTING

Jian Kang*, University of Michigan

Motivated by the needs of selecting important features from big neuroimaging data, we develop a new Bayesian feature screening approach in the generalized linear model (GLM) framework. We assign the conjugate priors on the coefficients and obtain the analytical form of the marginal posterior density function. Under some mild regularity conditions, we show that the marginal

posterior moments follow a mixture of normal distributions, one of which component is the standard normal distribution for unimportant variables. In light of this theoretical foundation, we develop a Bayesian variable screening algorithm for ultra-high dimensional data consisting of two steps: Step 1: compute a multivariate variable screening statistic based on marginal posterior moments; Step 2: perform the mixture model-based cluster analysis on screening statistics to identify the unimportant variables. Step 1 only requires a computational complexity on the linear order of the number of predictors and it is straightforward to be parallelized. It has a close connection with sure independent screening (SIS) statistics and high-dimensional ordinary least-squares projection (HOLP) methods. Step 2 is an extension of the local false discovery rate (FDR) analysis. We implement our method using massively parallel computing techniques based on the general-purpose computing on graphics processing units (GPGPU), leading to an ultra-fast variable screening procedure. Our simulation studies show that the proposed approach can perform variable screening on one million predictors within seconds and achieve higher selection accuracy compared with existing methods. We also illustrate our methods on an analysis of resting state functional magnetic resonance imaging (Rs-fMRI) data from the Autism Brain Imaging Data Exchange (ABIDE) study.

email: *jiankang@umich.edu*

THE BAYESIAN MULTIVARIATE REGRESSION FOR HIGH DIMENSIONAL LONGITUDINAL DATA WITH HEAVY-TAILED ERRORS

Viral V. Panchal*, Georgia Southern University

Daniel Linder, Georgia Southern University

Hani Samawi, Georgia Southern University

High-dimensional longitudinal data, also called “large p small n”, which consists of the situation when the number of measurements on subjects or sampling units is far greater than the size of the sample in the study. Similar to the popularity of longitudinal data in various biomedical and public health research, high-dimensional longitudinal data are also on the rise in bioinformatics, genomics, and public health research. These data exhibit heavy-tailed errors in frequent situations such as genomics, finance and more or contain outliers. Application of traditional ordinary least squares method for high dimensional longitudinal data will fail to produce valid estimates due to identifiability issues and specifically in heavy tails situation as it penalizes large deviations inappropriately. To address these issues, we present a method for variable selection and estimation based on the horseshoe prior for multivariate continuous outcomes with heavy-tailed errors. The proposed method is developed in a Bayesian setting and Gibbs sampler is derived to efficiently sample from the posterior distribution. We compare the method to standard estimation routines in a series of simulation examples as well as on a data set from a gene expression profiling experiment on T-cell activation.

email: *vp00187@georgiasouthern.edu*

51. GRAPHICAL MODELS

ESTIMATION OF HIGH-DIMENSIONAL GRAPHICAL MODELS USING REGULARIZED SCORE MATCHING

Lina Lin*, University of Washington

Mathias Drton, University of Washington

Ali Shojaie, University of Washington

Graphical models are widely used to model stochastic dependencies among large collections of variables. We introduce a new method of estimating undirected graphical independence graphs based on the score matching loss, introduced by Hyvarinen (2005), and subsequently extended in Hyvarinen (2007). The regularized score matching method we propose applies to settings with continuous observations and allows for computationally efficient treatment of possibly non-Gaussian exponential family models. In the well-explored Gaussian setting, regularized score matching avoids issues of assymetry that arrise when applying the technique of neighborhood selection, and compared to existing methods that directly yield symmetric estimates, the score matching approach has the advantage that the considered loss is quadratic and gives piecewise linear solution paths under l1-regularization. Under suitable irrepresentability conditions, we show that l1-regularized score matching is consistent for graph estimation in sparse high-dimensional settings. Through numerical experiments and an application to RNAseq data, we confirm that regularized score matching achieves state-of-the-art performance in the Gaussian case and provides a valuable tool for computationally efficient inference in non-Gaussian graphical models.

email: *linlina@uw.edu*

HIGH-DIMENSIONAL ROBUST PRECISION MATRIX ESTIMATION: CELLWISE CORRUPTION UNDER EPSILON-CONTAMINATION

Po-Ling Loh, University of Pennsylvania

Xin Lu Tan*, University of Pennsylvania

We analyze theoretical properties of robust estimators for inverse covariance matrices, when data are contaminated in a cellwise manner: each element of the data matrix is independently corrupted according to a certain proportion. Such contamination mechanisms may be used to model various phenomena in real-world scientific data, including measurement error in DNA microarray analysis and dropouts in sensor arrays. When data follow an uncontaminated

gaussian distribution, the graphical Lasso (GLasso) and CLIME algorithms are known to possess rigorous theoretical guarantees for the estimation of inverse covariance matrices in high dimensions; however, their performance may be compromised severely when data are contaminated by even a single outlier. The estimators we study are inspired by techniques in robust statistics and are constructed by plugging appropriately chosen robust covariance matrix estimators into GLasso and CLIME. We derive high-dimensional error bounds that reveal the interplay between the dimensionality of the problem and the degree of contamination permitted in the observed distribution, and also analyze the breakdown point of both estimators. Finally, we discuss implications of our work for gaussian graphical model estimation in the presence of contamination. Our results apply to arbitrary contaminating distributions and allow for a nonvanishing fraction of cellwise contamination.

email: *xtan@wharton.upenn.edu*

A NEW ORDINARY DIFFERENTIAL EQUATION MODEL FOR RECONSTRUCTION OF GENE REGULATORY NETWORK

Yaqun Wang*, Rutgers, The State University of New Jersey

Runze Li, The Pennsylvania State University

Rongling Wu, The Pennsylvania State University

Gene regulatory networks (GRN) play important roles in a complex living system. Using gene expression data to reconstruct GRN can enable one to gain new insights into the regulatory mechanisms underlying biological functions and phenotypic characteristics of an organism. Ordinary Differential Equation (ODE) modeling has been successfully used for inferrence of GRN based on time course gene expression, but this approach usually captures only instant regulation effects between genes. However, we know from biological knowledge that it takes time for regulator genes to have regulation impact on the targets. We address this issue by integrating the information of transcriptional time lags into ODE model to improve the accuracy of the GRN reconstruction. A procedure for GRN inferrence is developed with four steps including clustering genes, deciding potential regulators, detecting regulation effects and analyzing gene functions. A functional clustering approach based on Legendre Orthogonal Polynomial is applied in the first step for grouping genes according to the similarity of expression profiles over time. The model is also equipped with the power to jointly analyze data from multiple environments and, therefore, provides an unprecedented tool to help ones to understand a comprehensive picture of GRN. It has been well demonstrated by analyzing real data sets from a surgical research and through extensive simulation studies.

email: *yw505@sph.rutgers.edu*

DETECTING HIDDEN CHARACTERISTICS FOR NETWORK DATA WITHIN LATENT SPACE

Shiwen Shen*, University of South Carolina

Edsel Pena, University of South Carolina

Methods for detecting hidden community status for nodes in a network have been an important topic in the statistical network analysis. Most works have been done with respect to a snapshot of a network, while approaches for tracking movements for communities in a dynamic aspect have not been discussed frequently. We propose an idea of identifying the hidden characteristics for each node given the detected community status in an unknown hidden characteristics space, so that the problem of tracking communities can be transferred to making inference of the hidden characteristics of each node drift over time. We show how to find a p-dimensional hidden characteristics space by using the information of connections and covariates among nodes in the network; and how to generate a mapping to the space. Simulation results and an example using open sourced data are provided in detail.

email: *sshen@email.sc.edu*

STRUCTURED SPARSE MULTIPLE CO-INERTIA ANALYSIS WITH APPLICATIONS TO GENOMICS AND METABOLOMICS DATA

Eun Jeong Min*, Emory University Rollins School of Public Health

Qi Long, Emory University Rollins School of Public Health

Rapid advances in technology have led to the explosion of omics data in biomedical research. As a result, there has been an increasing interest in methods for integrative analysis of multiple types of omics data. Multiple co-inertia analysis is one of the statistical tools for assessing relationships and trends in more than two data types. While MCIA has been traditionally used in ecology area, more recently it has been used for integrative analysis of omics data. We propose a structured sparse multiple co-inertia analysis that incorporates biological information such as network information among genes or metabolites. Our proposed method is evaluated in simulations and illustrated in an application to integrative analysis of genomics and metabolomics data.

email: *ej.min@emory.edu*

NONPARAMETRIC MIXTURE OF GAUSSIAN GRAPHICAL MODELS, WITH APPLICATIONS IN BRAIN FUNCTIONAL CONNECTIVITY ESTIMATION

Kevin Haeseung Lee*, The Pennsylvania State University

Lingzhou Xue, The Pennsylvania State University

In many real-world applications, it is important to estimate heterogeneous conditional dependencies across the whole population because the observed data usually come from heterogeneous resources. In this work, we introduce a novel nonparametric mixture of Gaussian graphical models, which extends the methodology and applicability of Gaussian mixture models and time varying Gaussian graphical models. However, we need to address three significant challenges: high-dimensionality, non-convexity, and label switching. We propose a unified penalized likelihood scheme to effectively estimate both nonparametric functional parameters and heterogeneous graphical parameters, and further design a generalized effective EM algorithm to address three challenges simultaneously. We demonstrate our method in extensive simulation studies and also a real application to estimate brain functional connectivity from ADHD-200 Global Competition data, where two heterogeneous conditional dependencies are explained through profiling demographic variables and supported by existing scientific findings.

email: *khl119@psu.edu*

52. MULTIVARIATE METHODS

GROUPWISE ENVELOPE MODEL FOR EFFICIENT ESTIMATION AND RESPONSE VARIABLE SELECTION

Yeonhee Park*, University of Texas MD Anderson Cancer Center

Zhihua Su, University of Florida

The envelope model introduced by Cook et al. (2010) is a new paradigm that improves estimation efficiency in multivariate linear regression. In this article, we develop a couple of new envelope models that are more flexible and achieve more efficiency gains: The groupwise envelope model allows for distinct regression coefficients and error structures for different groups; and the sparse groupwise envelope model can identify response variables that are invariant to the changes in predictors. Theoretical properties of the proposed models are established. Numerical experiments show the effectiveness of the models in efficient estimation and variable selection.

email: *yeonhee@stat.ufl.edu*

CONSISTENT ESTIMATION IN PARTIALLY LINEAR MODELS WITH CORRELATED OBSERVATIONS

Liangdong Fan*, University of Kentucky

Cidambi Srinivasan, University of Kentucky

Richard Charnigo, University of Kentucky

Methods of estimating parametric and nonparametric components, as well as properties of the corresponding estimators, have been examined in partially linear models. These models are appealing due to their flexibility and wide range of practical applications. The compound estimator (Charnigo, Feng and Srinivasan, 2015) has been used to estimate the nonparametric component of such a model with multiple covariates, in conjunction with linear mixed modeling for the parametric component. These authors showed, under a strict orthogonality condition, that the parametric and nonparametric component estimators were (nearly) optimal, even in the presence of subject-specific random effects. Without orthogonality, iterative backfitting could be used to achieve convergence of both parametric and nonparametric estimators. However, the theoretical properties of those backfitted estimators are not established. Therefore, we now study both parametric and nonparametric estimators in a partially linear model with random effects, to extend the consistency results of Charnigo et al (2015) to a non-orthogonal design and those of Levine (2015) to correlated data. The random effects accommodate analysis of individuals on whom repeated measures are taken. We illustrate our estimators in a biomedical case study and assess their finite-sample performance in simulation studies.

email: *fanliangdong@uky.edu*

MULTIVARIATE MEAN ESTIMATION UNDER EFFICIENT SAMPLING DESIGNS

Daniel F. Linder, Georgia Southern University

Haresh D. Rochani*, Georgia Southern University

Hani M. Samawi, Georgia Southern University

Viral V. Panchal, Georgia Southern University

In many studies, the researchers attempt to describe the relationship between more than two outcome (or response) variables with its determinants (covariates). In this paper, we present an efficient procedure based on ranked set sampling to estimate and perform the hypothesis testing on a multivariate outcome mean. The method is based on ranking on an auxiliary covariate, which is assumed to be correlated with the multivariate response, in order to improve the efficiency of the estimation. We show that the proposed estimator developed under this sampling scheme is unbiased, has smaller variance in the multivariate sense, and is asymptotically Gaussian. A bootstrap routine is developed in the statistical software R to perform the inference when the sample size is small. We also extend the regression estimators based on ranked set sampling to multivariate regression. We use a simulation study to investigate the performance of the method under known conditions and apply the method to the biomarker data collected in China Health and Nutrition Survey (CHNS 2009) data.

email: *hr00178@georgiasouthern.edu*

GLOBAL RANK TESTS FOR MULTIPLE ORDINAL AND FAILURE OUTCOMES

Ritesh Ramchandani*, Harvard School of Public Health
David A. Schoenfeld, Massachusetts General Hospital
Dianne M. Finkelstein, Massachusetts General Hospital

Treatments evaluated in clinical trials may be expected to affect the patient on many dimensions. For example, treatments for a neurological disease such as ALS are often intended to impact several dimensions of neurological function, as well as survival. The assessment of treatment on the basis of multiple outcomes is a challenging problem, both in terms of selecting a valid test and interpreting the results. Several nonparametric global tests have been proposed (e.g. O'Brien), and we generalize these into a broader, flexible framework using U-statistics. The generalized test is based on a simple scoring mechanism applied to each pair of subjects for each endpoint. The pairwise scores are then reduced to a summary score, and a rank-sum test is applied to the summary scores. For certain tests, we establish optimal outcome weighting schemes based on power and relative importance of the endpoints. Since the optimal weights depend on alternatives that are often unknown in practice, we also propose an adaptive weighting method, and evaluate the type 1 error and power in simulation studies. The methods are applied to analyze the impact of a treatment on neurological function rating and mortality in an ALS clinical trial.

email: *rir072@mail.harvard.edu*

ENVELOPE MODELS FOR EFFICIENT MULTIVARIATE BINARY REGRESSION

Emil A. Cornea*, University of North Carolina, Chapel Hill
Joseph G. Ibrahim, University of North Carolina, Chapel Hill
Hongtu Zhu, University of North Carolina, Chapel Hill

The envelope models constitute a new framework to address estimation and prediction in multivariate analysis. We develop such an envelope model for analyzing correlated multivariate binary data such as repeated-measures data or multiple-indicators with measures of some underlying characteristic using number of outcomes reduction techniques, a new version of the classical multivariate logistic regression model. It can serve as a mean to reinterpret and improve efficiency for a range of multivariate statistical methods. Our approach is based on a copula model of underlying latent threshold random variables. It yields likelihood-based models for marginal fixed effects estimation and interpretation in the analysis of correlated binary data. The asymptotic distribution and the consistency of its maximum likelihood estimators are established. The MLE for the proposed model can be substantially less variable then the usual MLE. The maximization is carried out via the EM algorithm by treating the latent random variables as missing data. We illustrate our approach on simulated and real datasets. The model is applied to analyze data from an environmental study involving dyspnea in cotton workers.

email: *ecornea@bios.unc.edu*

A GEOMETRIC PERSPECTIVE ON THE POWERS OF PRINCIPAL COMPONENT ASSOCIATION TESTS IN MULTIPLE PHENOTYPE STUDIES

Zhonghua Liu*, Harvard University
Xihong Lin, Harvard University

Principal component analysis (PCA) has been commonly used for analyzing multiple phenotypes in genetic association studies. Previous studies empirically found that the top few principal components (PCs) might not be powerful to detect the underlying causal genetic variants. However, little theoretical work has been done regarding when PCA is powerful and when it is not. In this paper, we use eigen-analysis and theoretical power analysis from a geometric perspective to investigate why the top PCs might be powerless. To increase power, we propose linear, nonlinear and adaptive combination of PCs. Our methods are based on GWAS summary statistics so that individual level data are not necessary. We further conduct extensive simulation studies to investigate how the dimension of phenotypes, signal sparsity, effect homogeneity and the correlation structures influence the powers of our methods. The simulation results show that our methods all maintain valid type I error rates and their empirical powers are consistent with the theoretical analysis. We further apply our methods to a global lipids level genome-wide association study data set and identify hundreds of novel genetic variants that were missed by the original single-trait analysis approaches. We also develop an R package freely available for public uses.

email: *zliu@mail.harvard.edu*

MULTILEVEL MATRIX-VARIATE ANALYSIS AND ITS APPLICATION TO LONG-TERM REMOTE PATIENT MONITORING

Lei Huang*, Johns Hopkins University
Tamara Harris, National Institute of Aging, National Institutes of Health
Mathew Maurer, Columbia University Medical Center
Philip Green, Columbia University Medical Center
Andrada Ivanescu, Montclair State University
Vadim Zipunnikov, Johns Hopkins University

The number of studies where the primary measurement is a matrix is exploding. In response to this, we propose a statistical framework for modeling populations of repeated matrix-variate measurements. We use a linear mixed effect model to account for the multilevel design, while the 2D structure is handled via normal matrix-variate distribution. Row- and column-specific covariance operators are estimated by the method of moments and diagonalized to achieve dimension reduction. The computational feasibility and performance of the approach is shown in extensive simulation studies. The method is motivated by and applied to a study that remotedly monitored physical activity of individuals diagnosed with congestive heart failure (CHF). Participants wore an accelerometer that continuously recorded physical activity over a 3- to 10-month period. Two primary goals of the study were: 1) to quantify and model the long-term patterns of physical activity in individuals with CHF; and 2) evaluate the possibility of predicting adverse health effects via continuous activity monitoring.

email: *huangracer@gmail.com*

53. ORAL POSTERS: CLINICAL TRIALS

53a. INVITED ORAL POSTER: ADAPTIVE PLATFORM TRIALS: THE FUTURE OF CLINICAL RESEARCH

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

Platform trials consider multiple therapies for a particular disease. When the therapies are drugs or combinations of drugs, they may be owned by the same company. Increasingly, however, and spurred by government agencies and philanthropy, companies are collaborating. Some collaborations are simple unions of clinical trials that have separate designs but are unified under a single master protocol. More relevant for the future are trials with many arms that include combinations of drugs from different companies with explicit and implicit comparisons of therapies across companies, with adaption randomization to identify how to deliver better therapy to patients by disease subtype. I will give examples in cancer, Alzheimer's disease, and community-acquired pneumonia.

email: *don@berryconsultants.net*

53b. INVITED ORAL POSTER: STATISTICAL DESIGN AND ISSUES IN A SCIENTIFIC BREAKTHROUGH TRIAL FOR HIV PREVENTION

Ying Qing Chen*, Fred Hutchinson Cancer Research Center

The HIV Prevention Trial Network 052 Study is a Phase III, controlled, randomized clinical trial to assess the effectiveness of immediate versus delayed antiretroviral therapy strategies on sexual transmission of HIV-1 (Cohen, et al., 2011, New England Journal of Medicine). It was hailed by the Science Magazine as the Breakthrough of the Year for 2011 (Alberts, 2011, Science). In this poster, we will highlight the statistical design and issues that underlie this successful trial in HIV Treatment-as-Prevention, and summarize the lessons that we have learned for future research.

email: *yqchen@fredhutch.org*

53c. ESTIMATION OF DOSAGE FREQUENCY OF PRE-EXPOSURE PROPHYLAXIS NEEDED TO PROTECT AGAINST HIV INFECTION

Claire F. Ruberman*, Johns Hopkins University

WITHDRAWN

We apply Targeted Maximum Likelihood Estimation (TMLE) methodology to data from a case-cohort analysis within the Partners Pre-Exposure Prophylaxis (PrEP) Tenfovoir clinical trial to infer the marginal effect of plasma Tenofovir (TNF) levels on reducing the risk of seroconversion. We explore the effect of having quantifiable levels of TNF, accounting for adherence, to estimate an interval for the relative risk reduction. We model adherence by fitting a marginal structural model within TMLE and map the trajectory of risk by coverage when stratified by plasma concentration. We apply Targeted Maximum Likelihood Estimation (TMLE) methodology to data from a case-cohort analysis within the Partners Pre-Exposure Prophylaxis (PrEP) Tenfovoir clinical trial to infer the marginal effect of plasma Tenofovir (TNF) levels on reducing the risk of seroconversion. We explore the effect of having quantifiable levels of TNF, accounting for adherence, to estimate an interval for the relative risk reduction. We model adherence by fitting a marginal structural model within TMLE and map the trajectory of risk by coverage when stratified by plasma concentration. For quantifiable TNF, the curve is nearly flat and for below quantifiable TNF, risk decreases with pill count. We explore how much the protective effect of Tenofovir is mediated through plasma TNF concentration, employing TMLE to control for confounding in estimating the controlled direct. When the TNF concentration is fixed, the estimated risk reduction of seroconversion of setting coverage to high is negligible, suggesting the protective effect of Tenofovir is almost fully mediated through plasma TNF. We extend our methods to three additional PrEP studies: the CDC Bangkok Tenofovir study of injection drug users, the iPrEx study of men and transgender women, and the VOICE study of heterosexual women.

email: *claireruberman@gmail.com*

53d. A MIXTURE OF MIXED LOGISTIC REGRESSION MODEL FOR DYNAMIC TREATMENT REGIME WITH APPLICATION TO PROSTATE CANCER TRIAL

Bing Yu*, Purdue University

Bruce Craig, Purdue University

Yu Zhu, Purdue University

In clinical trials, patients respond differently to the same treatment due to patient heterogeneity and chance variation. In this work, we focus on one type of heterogeneity, which is the existence of distinct subgroups of patients that respond similarly to a set of different treatments. We propose to use the mixture of mixed logistic regression model to estimate subgroup proportions and subgroup-specific treatment effects. Optimal dynamic treatment regime can be determined based on these results. Several algorithms of estimation are provided as well as consideration of the identifiability conditions for the model parameters. The proposed model is applied to both a sequential multiple assignment randomized trial (SMART) and a crossover design. Simulation studies are performed to demonstrate the effectiveness of the proposed methods. We further apply the proposed model and methods to analyze a prostate cancer trial data and compare different dynamic treatment regimes.

email: *yu245@purdue.edu*

53e. UNDERSTANDING THE OPERATING CHARACTERISTICS OF DIFFERENT BAYESIAN ADAPTIVE ALLOCATIONS IN TWO ARM CONFIRMATORY TRIAL WITH A DICHOTOMOUS OUTCOME

Yunyun Jiang*, Medical University of South Carolina

Wenle Zhao, Medical University of South Carolina

Valerie L. Durkalski, Medical University of South Carolina

Background. Bayesian response adaptive allocation algorithms are growing in popularity in the clinical trial arena. Various adaptive allocation formulas are available however differences between the methods have rarely been examined. Materials and methods. A simulation study is conducted to compare the performance of three adaptive allocation algorithms: square root transformation AR(1/2); proportion of concurrent sample size AR(n/2N) (Thall and Wathen 2007); and, inclusion of a variance component AR(1/2, ?) (Berry et al. 2010). Allocation targets were updated every 50 enrollments as well as per accumulating patient. The operating characteristics are compared to fixed equal allocation under frequentist and Bayesian group sequential designs. Results. Compared to equal allocation, the three adaptive allocations reduce the proportion of patients receiving the inferior treatment, at the cost of power. AR (n/2N) creates the smallest treatment imbalance, and has the highest power. AR (1/2, ?) exhibits robustness and generates more stable allocation probabilities over the course of the trial. The bias adjustment improves the precision of the treatment estimate for all three adaptive allocations. Conclusion. Bayesian adaptive allocation is a viable alternative to fixed allocation schemes in certain settings. However, the per-accumulating recruitment update should be avoided as it tends to reduce any ethical benefit.

email: *jiany@musc.edu*

53f. USING EVENT COUNTS IN PHASE I CLINICAL TRIALS

Daniel G. Muenz*, University of Michigan

Thomas M. Braun, University of Michigan

Jeremy M. G. Taylor, University of Michigan

Phase I clinical trials in oncology that seek to identify the MTD -- the highest dose with acceptable toxicity -- typically dichotomize the data: a patient either has a dose-limiting toxicity (DLT) or not. We propose a simple Poisson-binomial model that works with a richer set of data for each patient that includes both the number of DLTs as well as the number of mild toxicities (non-DLTs). As a result, our model provides deeper information about the toxicity profile of each dose level under investigation, as compared to the standard continual reassessment method (CRM) models. The trade-off is that our model has more parameters to estimate than the CRM. However, the increased complexity of our model can be mitigated in a Bayesian framework by appropriate tuning of prior distributions. Specifically, we can calibrate the prior variances used in each method so that both models lead to similar prior probabilities of DLT at each dose. We present simulation results showing that our model performs similarly to the CRM in adaptive dose-finding trials, with slight gains in identification of the MTD when there is modest variation in the total number of toxicities (both DLT and non-DLT) among the doses.

email: *dmuenz@umich.edu*

53g. META-ANALYSIS OF CLINICAL TRIALS WITH SPARSE A BINARY OUTCOMES USING ZERO-INFLATED BINOMIAL (ZIB) MODELS

Cheng Dong*, University of Missouri

Yueqin Zhao, U.S. Food and Drug Administration

Ram Tiwari, U.S. Food and Drug Administration

Recently, meta-analysis has been widely used in clinical studies for evaluating the safety or efficacy of the drug. When dealing with the rare events data, a large number of studies have no event of interest. In order to estimate the pooling odds ratio, people usually exclude such studies or apply continuity correction on zero-event arms. However, excluding ZTE studies may lead to invalid or inefficient inference and different continuity corrections may result in different conclusions. In this paper, we apply zero-inflated Binomial (ZIB) model on clinical trial data with sparse Binary outcomes and evaluate the performance of the proposed ZIB model via simulated data. The proposed ZIB model outperforms the Binomial model, Mantel-Haenszel and Peto methods and it does a good job in estimating odds ratio. We also illustrate the proposed ZIB model with Rosiglitazone cardiovascular deaths data with 48 trials.

email: *cd5w4@mail.missouri.edu*

53h. RESPONSE ADAPTIVE RANDOMIZATION USING SURROGATE AND PRIMARY ENDPOINTS

Hui Wang*, Virginia Commonwealth University
Nitai Mukhopadhyay, Virginia Commonwealth University

In recent years, adaptive designs in clinical trials have been attractive due to its efficiency and flexibility. Response adaptive randomization procedures in phase III clinical trials are proposed to appeal ethical concerns by skewing the probability of patient assignments based on the responses obtained thus far, so that more patients will be assigned to a superior treatment group. General response-adaptive randomizations usually assume that the primary endpoint can be obtained quickly after the treatment. However, in real clinical trials, one may need to take a relatively long time to observe the primary endpoints. Thus, a new response-adaptive randomization method will be proposed which incorporates one or more surrogate endpoints that are correlated with the primary endpoint. Our method utilizes surrogate and primary endpoints at the same time for clinical trials with continuous responses. The parameter estimates will be based on the conditional distribution of primary endpoint given one or more surrogate endpoints through a Bayesian model. Then these parameters will be plugged into the desired target allocation to obtain the optimal proportion to one treatment. Finally, we will use these sequentially estimated proportions based on Doubly Adaptive Biased Coin Design (DBCD) rule to skew the allocation.

email: *wanghui33366@gmail.com*

53i. EFFICIENT DOUBLE ROBUST ESTIMATION FOR TWO-STAGE DYNAMIC TREATMENT REGIMES

Andrew S. Topp*, University of Pittsburgh
Geoff S. Johnson, University of Pittsburgh
Abdus S. Wahed, University of Pittsburgh

Certain conditions and illnesses may necessitate multiple stages of treatment and thus require unique study designs to compare the efficacy of these interventions. Such studies evaluate dynamic treatment regimes (DTRs) that are characterized by two or more stages of treatment punctuated by decision points where a patient's information up to that point informs their treatment assignment for the next stage. There are various methods of estimating the effect of DTRs from sequential designs. A doubly robust estimator utilizes both modeling of the outcome and weighting based on the modeled probability of receiving treatment. This estimator gives a consistent and unbiased estimate of the desired population parameter under the condition that at least one of those models is correct. This paper describes a method of building doubly robust estimators of treatment effect of different regimes using an efficient means of estimating outcome model coefficients. This new and more efficient doubly robust estimator is compared to existing doubly robust estimators as well as g-computation and inverse probability weighted estimators in a simulation study. This new estimator is then used to estimate regime mean outcomes in the STAR*D Anti-depression Treatment Trial.

email: *AST25@pitt.edu*

53j. CHOOSING COVARIATES FOR ADJUSTMENT IN NON-INFERIORITY TRIALS BASED ON INFLUENCE AND DISPARITY

Katherine S. Nicholas*, Medical University of South Carolina
Viswanathan Ramakrishnan, Medical University of South Carolina
Valerie L. Durkalski-Mauldin, Medical University of South Carolina

It has been shown that the type I error is inflated when important covariates are excluded from a non-inferiority analysis (Nicholas et al, 2014). Traditionally, whether or not to adjust for a covariate in a model is based solely on statistical significance or some other criteria that relates to the magnitude of the effect. In addition, one may also choose to perform tests of baseline imbalance. However, several authors suggest that these aspects should be considered simultaneously. For example, Canner et. al. (1991) developed a statistic to determine the relative importance of including a covariate in a model based on both its effect on the outcome (which he calls influence) and its association with treatment (which he calls disparity). Although Canner et. al.'s approach assumed no treatment effect, Beach et. al. (1989) extended this to non-zero treatment effects in the context of linear regression. The current research seeks to combine the methods of Canner et. al for binary outcomes with the methods of Beach et. al for non-zero treatment effect in order to quantify the relative importance of including covariates in a non-inferiority study with a binary outcome. Theoretical results are presented and applied via simulation, followed by practical application.

email: *nicholk@musc.edu*

53k. BAYESIAN MODELING AND PREDICTION OF ACCRUAL USING GAUSSIAN PROCESS

Yi Deng*, Emory University
Qing He, Emory University
Qi Long, Emory University

Phase III clinical trials usually require adequate sample size. In practice, it is more reasonable and possible to enroll patients from more than one regions such that the drug development can follow a scheduled timeframe. Therefore, proper monitoring and prediction of patient accrual is essential in multi-regional trials. Standard methods used to predict the patient accrual are over-simplified by assuming a constant or piecewise constant accrual rate. We propose and evaluate a time and region dependent Bayesian model of accrual rate based on Gaussian process. For some large regions, there are usually sub accrual units, i.e. sites. We further propose a hierarchical Bayesian model that can take the additional spatial information into consideration. In numerical studies, we show that our proposed models out-perform other models.

email: *ydeng26@emory.edu*

53I. STEPPED WEDGE CLUSTER RANDOMIZED CONTROLLED TRIALS WITH TWO LAYERS OF CLUSTERING: DESIGNS AND COMPARISONS OF POWER

Ranran Dong*, The Ohio State University

Abigail Shoben, The Ohio State University

The stepped wedge cluster randomized trial (SW-CRT) is a family of cluster randomized trials in which groups of participants rather than individual participants are randomly allocated. In this paper, we propose several novel stepped wedge cluster designs for data with multiple layers of clustering. We focus on cross-sectional designs in which different participants from each cluster are observed at each step. Given our designs and an existing design, we study the efficiency of the treatment parameter estimator at different intra-cluster correlation (ICC) values. Given data with two layers of clustering, the designs discussed in this paper differ in the allocation of steps where units from the same cluster are exposed to the intervention. In design 1 (the existing design), all units in the same cluster transfer to the intervention at a single step. In designs 2 and 3, units from the same cluster complete the transition to the intervention within two adjacent and nonadjacent steps, respectively. In design 4, units in the same cluster are allocated to the intervention at each step. According to the results of our simulation studies, design 4 yields the most efficient estimator of the treatment parameter, since it maximizes the number of between-unit within-cluster comparisons.

email: *dong.237@osu.edu*

54. NEW STATISTICAL METHODS FOR IMAGING GENETICS

TESTING FOR ASSOCIATION BETWEEN GENETIC VARIANTS AND BRAIN NETWORKS

Junghi Kim, University of Minnesota

Wei Pan*, University of Minnesota

It is now believed that many mental and psychiatric disorders, such as Alzheimer's disease, are caused by disrupted brain functional and/or structural networks. We propose using brain networks as an intermediate phenotype for Alzheimer's disease and conducting a genome-wide association scan on single nucleotide polymorphisms (SNPs). In addition, to boost power, we also propose both gene- and pathway-based analyses. We will use the ADNI data as an example.

email: *weip@biostat.umn.edu*

INTEGRATING GENOMIC AND IMAGING DATA: AN ATOMIC APPROACH

Debashis Ghosh*, Colorado School of Public Health

With the advent of private-public partnerships like the Alzheimer's Disease Neuroimaging Initiative (ADNI) and the Parkinson's Progression Markers Initiative (PPMI), there is a lot of scientific and medical consideration being given to the joint analysis of genetic and imaging data in large-scale studies. Given that this represents a merging of two big-data sources, developing statistical methods that are both computationally scalable as well as having mathematically knowable properties is of paramount importance. In this talk, we describe an approach to the analysis of such data that we term atomic. Doing so will unify many of the sparse regression methods available in the literature as well as lead to new methods that satisfy the two criteria mentioned above. In addition, the concept of an atom is fairly generic and can refer to many objects of inferential interest, including individual genes, anatomical structures in the brain and networks components. We will illustrate our methods using both simulated and real data.

email: *debashis.ghosh@ucdenver.edu*

JOINT MEDIATION ANALYSIS OF IMAGING AND GENETIC DATA IN GENETIC ASSOCIATION STUDIES OF COMPLEX DISEASES

Hongtu Zhu*, University of North Carolina, Chapel Hill

Many large-scale imaging genetic studies have been conducted to collect a rich set of imaging, genetic, and clinical data to detect putative genes for complexly inherited neuropsychiatric and neu-

rodegenerative disorders. The aim of this paper is to develop a joint mediation association analysis of imaging and genetic data (MASIG) framework to efficiently carry out genetic association studies of complex diseases. MASIG is an extension of standard mediation framework to the causal mediation analysis framework with the imaging measures as a potential mediator. We propose a variance component test for the total effect of SNPs and imaging measures on disease risk. Finally, we have successfully applied MASIG to a large-scale imaging genetic data analysis of ADNI data with 708 subjects. Our MASIG may be a valuable statistical toolbox for large-scale imaging genetic analysis as the field is rapidly advancing with ultra-high-resolution imaging and whole-genome sequencing.

email: *hzhu@bios.unc.edu*

HERITABILITY-BASED PRIORITIZATION OF STRUCTURAL NEUROIMAGING PHENOTYPES

Tian Ge*, Harvard Medical School

Thomas E. Nichols, University of Warwick

Martin Reuter, Harvard Medical School

Anderson M. Winkler, University of Oxford

Avram J. Holmes, Yale University

Phil H. Lee, Harvard Medical School

Joshua L. Roffman, Harvard Medical School

Randy L. Buckner, Harvard University

Jordan W. Smoller, Harvard University

Mert R. Sabuncu, Harvard Medical School

Measurements from structural brain magnetic resonance imaging (MRI) scans have been increasingly analyzed as intermediate phenotypes to bridge the gap between clinical features and genetic variation. To date, most imaging genetics studies have focused on regional and univariate (scalar) neuroimaging phenotypes, such as the volume or average cortical thickness of a brain structure. While these coarse quantifications of structural integrity have yielded important discoveries, they fail to exploit the richness of brain imaging data and to capture the complex geometry of neuroanatomy. Here we present statistical methods that enable the heritability analysis of massive phenotypes (e.g., voxel-level neuroimaging measurements) and multidimensional phenotypes (e.g., neuroanatomical shape measurements), using genome-wide single nucleotide polymorphism (SNP) data from unrelated individuals. As a demonstration of application, we apply our methods to the Harvard/Massachusetts General Hospital (MGH) Brain Genomics Superstruct Project (GSP), a unique and large-scale neuroimaging genetics study, to construct high-resolution surface maps of the heritability of morphological measurements, and to conduct the first comprehensive heritability analysis of the shape of anatomical structures spanning the human brain. The heritability profile of structural neuroimaging phenotypes can be useful for the identification of imaging endophenotypes related to brain disorders.

email: *tge1@mgh.harvard.edu*

55. RECENT ADVANCES IN ADAPTIVE MODEL-BASED DESIGN OF CLINICAL TRIALS

ADAPTIVE DOSE ESCALATION METHODS IN PHASE I ONCOLOGY TRIALS: A CASE STUDY

Inna Perevozskaya*, Pfizer Inc.

Roberto Bugarini, Pfizer Inc.

M**ani Lakshminarayanan**, Pfizer Inc.

Last decade has seen an increased use of innovative adaptive designs in phase 1 oncology trials. Modified Toxicity Probability Interval (mTPI) and Continued Reassessment Method (mCRM) are most commonly used such methods. Both algorithms learn about dose-toxicity using interim data but they handle the data differently: mTPI models probability of toxicity at each dose separately, while CRM relies on parametric dose-toxicity model pulling information across doses. In this case study, we have compared mTPI, mCRM and 3+3, by simulation, to quantify the trade-off between precision of MTD estimation and design costs. The mTPI method offered substantial gain in MTD estimation over classical 3+3 design while maintaining transparency of all dose-escalation decisions and ability to pre-specify them before trial starts. On the other hand, mCRM was more complex and less transparent than both 3+3 and mTPI. Despite its additional gains over mTPI in correct MTD determination, the mCRM was not chosen as the final design. Additional benefits of mCRM were modest and not sufficient to outweigh added complexity and larger sample size/study duration. The study is on-going at the time this abstract is written; we plan to share study results and lessons learned, if available.

email: *inna.perevozskaya@pfizer.com*

EARLY-PHASE DESIGN FOR A COMBINATION OF TARGETED THERAPIES IN ONCOLOGY

Nolan A. Wages*, University of Virginia

Craig A. Portell, University of Virginia

Gina R. Petroni, University of Virginia

In early-phase oncology trials, treating patients with novel targeted therapies is becoming increasingly popular. For cytotoxic agents, the fundamental assumption driving the design is that both toxicity and efficacy increase monotonically with dose. Therefore, the primary objective has been to find the maximum tolerated dose (MTD), with the assumption that the highest safe dose also provides the most promising outlook for efficacious benefit. By contrast, targeted agents may exhibit non-monotone dose-response patterns, meaning that higher doses may not correlate with greater efficacy. Innovative dose-finding strategies are needed to establish the safety and efficacy of these novel therapies. This talk describes the implementation of an adaptive early-phase method for identifying the optimal combination of two oral targeted inhibitors. Operating characteristics of the design are demonstrated under various possible true scenarios via simulation studies. The results demonstrate the method's ability to effectively recommend optimal combinations in a high percentage of trials with manageable sample sizes. Overall performance indicates that the design is a practical early-phase design for use with combined targeted agents.

email: *nwages@virginia.edu*

INFERENCE IN EXPERIMENTS WITH PILOT DATA

Nancy Flournoy*, University of Missouri

Multi-stage experiments are typical in clinical trials, biology, biochemistry, and other fields. Even most basic experimental designs are functions of unknown parameters, and prior information is needed to plan an efficient study. Suppose pilot study data will be used to determine how to allocate subjects to treatments in a larger study. We discuss the dependencies in the combined data set, and options for inference when such dependencies are present. Indeed, distributional complications typically disappear if both stages have large sample sizes. Of more interest is the finite sample case. To illuminate the issues, we consider non-linear regression models with normal errors. The features of several approximate distributions for maximum likelihood estimates are compared. Because maximum likelihood estimates are often skewed when sample sizes are small, we also discuss the potential role of the bootstrap in this setting.

email: *flournoyn@missouri.edu*

56. STATISTICAL METHODS FOR ACCELEROMETRY DATA: PHYSICAL ACTIVITY AND HEALTH OUTCOMES

QUANTIFYING PHYSICAL ACTIVITY IN MID-TO-LATE LIFE

Jennifer A. Schrack*, Johns Hopkins University

Vadim Zipunnikov, Johns Hopkins University

Luo Xiao, North Carolina State University

Ciprian Crainiceanu, Johns Hopkins University

Luigi Ferrucci, National Institute on Aging, National Institutes of Health

It is well established that physical activity has beneficial effects on health and longevity, yet the majority of adults remain relatively sedentary. Despite decades of promoting physical activity across the life course, remarkably little is known about how daily physical activity quantities, patterns, and trends vary with age. Moreover, because of difficulties with measuring light intensity physical activity, very little is known about the health benefits of "lifestyle" activities. Advances in physical activity monitoring through accelerometers provide clinicians and researchers with unprecedented opportunities to further understanding of the health benefits of all levels of physical activity, and to detail trajectories of physical activity according to health and functional status. Data from the Baltimore Longitudinal Study of Aging (BLSA) will be presented that depict detailed changes in quantities, patterns, and trends of objectively measured daily physical activity in mid-to-late life, and emphasize the methodological challenges associated with defining factors contributing to these changes.

email: *jschrac1@jhu.edu*

MULTILEVEL MODELS FOR ANALYSIS OF ACTIGRAPHY DATA

Vadim Zipunnikov*, Johns Hopkins University

Jeff Goldsmith, Columbia University

Junrui Di, Johns Hopkins University

Andrew Leroux, Johns Hopkins University

Jacek Urbanek, Johns Hopkins University

The first part of my talk will provide a quick review of the strengths and limitations of current analytical approaches for modeling physical activity data. In the second part, I will talk about analysis of physical activity data collected on 13K+ subjects in the National Health and Nutrition Examination Survey (NHANES), a nationally representative sample of the US population. I will present recent multilevel functional data approaches

that separate and quantify the systematic and random circadian patterns of physical activity, model them as functions of age, gender, and dominant comorbidities and demonstrate that these patterns are powerful predictors of mortality.

email: *vadim.zipunnikov@gmail.com*

THREE-PART JOINT MODELING METHODS FOR COMPLEX FUNCTIONAL DATA IN PHYSICAL ACTIVITY STUDIES

Haocheng Li*, University of Calgary

John Staudenmayer, University of Massachusetts

Tianying Wang, Texas A&M University

Raymond J. Carroll, Texas A&M University

We take a functional data approach to longitudinal studies with complex bivariate outcomes. One response is obtained in continuous proportions with excess zeros and ones. The other outcome is a continuous variable featured by excess zeros and skewness. A three-part functional data joint modeling approach is introduced. The first part is a continuation-ratio model to postulate the ordinal features for proportional response to be 0, (0,1) or 1. The second part is to model the proportions when they are in interval (0,1). The last component specifies the skewed continuous measurements by Box-Cox transformations when the proportions are at (0,1) or 1. In this three-part model, the regression structures are specified as smooth curves measured at various time-points with random effects that have a correlation structure. The smoothed random curves for each variable are summarized using a few important principal components, and the association of the three longitudinal components is modeled through the association of the principal component scores. The difficulties in handling the ordinal and proportional variables are solved by using a quasilikelihood type approximation. We develop an efficient algorithm to fit the model, which involves the selection of the number of principal components. The method is applied to physical activity data, and is evaluated empirically by a simulation study.

email: *haocheng.li@ucalgary.ca*

ACCELEROMETERS, PHYSICAL ACTIVITY, AND CONDITIONAL RANDOM FIELDS

Evan Ray, University of Massachusetts, Amherst

John W Staudenmayer*, University of Massachusetts, Amherst

We consider the problem of recognizing aspects of physical activity from body worn accelerometers. The accelerometers provide vector valued observations at each time point, and the statistical problem is to classify the concurrent activity that the wearer is doing. We compare several approaches: hidden Markov models, conditional random fields, and static models. We apply our methods to simulated data and to several real datasets where what the wearers are actually doing is known from direct observation.

email: *jstauden@math.umass.edu*

57. STATISTICAL METHODS IN HIV/AIDS

INTEGRATION OF DYNAMIC GENE REGULATORY NETWORKS FOR HIV INFECTION IN THE BIG DATA ERA

Hulin Wu*, University of Texas Health Science Center, Houston

It is important to understand HIV pathogenesis at gene, protein, and cell levels in order to eradicate HIV and cure AIDS. More and more data sets at all these levels are available at public databases owing to the data sharing policy. In particular, many time course high-throughput gene expression data for HIV infection under different experimental conditions are available in the GEO database. It is challenging to identify all these data and integrate these heterogeneous data sets to extract meaningful information. In this work, we employ the text mining and ontology techniques to identify relevant data sets and apply a sophisticated analytic pipeline to the identified data sets to reconstruct the dynamic gene regulatory networks using high-dimensional ordinary differential equation approaches. We integrate the results from multiple data sets for HIV infection under different experimental settings. In this talk, I will present the analytic pipeline and biological findings. I will also discuss the extension of the methodology to address other scientific questions.

email: *Hulin.Wu@uth.tmc.edu*

ESTIMATING THE ASSOCIATION OF BIVARIATE SURVIVAL DATA THROUGH COPULA MODELS: AN APPLICATION TO A STUDY OF AIDS-RELATED NON-HODGKIN'S LYMPHOMA IN EAST AFRICA

Pingfu Fu*, Case Western Reserve University

Xiaozhen Han, Case Western Reserve University

Scot Remick, Mary Babb Randolph Cancer Center

Forty-nine patients with AIDS-related non-Hodgkin's lymphoma in East Africa were treated with lomustine, VP-16, cyclophosphamide and procarbazine. There is a strong interest in the association of overall survival (OS) and progression-free survival (PFS) in the area of biomarkers and surrogate endpoints research. We hypothesize there is a positive association between OS and PFS in this specific disease setting. Due to censoring, the estimation and inference of the association parameter of bivariate survival data based on methods developed recently, namely, a semi-parametric normal copula-based approach, in particular Spearman correlation coefficient, as the dependence of such times is assumed monotonic, is

used. A simulation study was conducted to explore which copula is optimal for such task under various scenarios. Based the simulation study, the correlation of the two times with censoring for the clinical study, rs = 0.91 (95% CI: 0.02 - 0.46, $p < 0.0001$), shows that there was a strong positive association between OS and PFS.

email: *pxf16@case.edu*

A DIRICHLET PROCESS MIXTURE MODEL FOR NON-IGNORABLE DROPOUT

Camille Marie Moore*, University of Colorado Denver

Samantha MaWhinney, University of Colorado Denver

Nichole E. Carlson, University of Colorado Denver

Longitudinal cohorts are a valuable resource for studying HIV disease progression, however dropout is common in these studies, with subjects often failing to return for visits due to disease progression, loss to follow-up, or death. When dropout depends on unobserved outcomes, data is missing not at random and results from standard longitudinal data analyses can be biased towards subjects that remain on study, who likely have more favorable outcomes. Several methods have been proposed to adjust for non-ignorable dropout; however, many of these approaches rely on parametric assumptions about the distribution of dropout times and the relationship between the outcome and dropout time. We propose a Bayesian semi-parametric Dirichlet process mixture model to relax these assumptions and provide more accurate inference when parametric assumptions are violated. Results from simulation studies as well as an application to a publicly available longitudinal HIV cohort study database are presented.

email: *camille.moore@ucdenver.edu*

A STATE SPACE FRAMEWORK FOR PATIENT-LEVEL MODELING OF THE HIV CARE CASCADE USING LONGITUDINAL COHORT DATA

Hana Lee*, Brown University

Joseph W. Hogan, Brown University

Becky L. Genberg, Brown University

Paula Braitstein, Indiana University

The HIV care cascade is a conceptual model used to describe the steps that patients with HIV must take to achieve viral suppression. Policy makers and clinical researchers are interested in rates of progression through each stage of the cascade, and in individual characteristics that predict retention or gaps in care. Analytic methods used to address these questions typically rely on targeted aspects of the cascade using data from a single cohort, or on simulation of compartment models of the full cascade that rely on heterogeneous inputs derived from descriptive statistics or published studies. To capture the advantages associated with using a single cohort with the complexity of a compartmental model, we have developed a state-space modeling (SSM) approach that provides a unified statistical analysis of a large cohort of patients as they pass through the cascade of HIV care. The SSM is parameterized in terms of probability of transition from one state to another, but also allows regression-based estimation of (time-varying) covariate effects associated with transitions between and within states to capture comprehensive patient-level behaviors along the stages of the cascade while accounting for temporal variations and within-subject correlations. A simple version of the model is illustrated using data from the Academic Model Providing Access to Healthcare (AMPATH), HIV care and treatment program in western Kenya.

email: *hana_lee@brown.edu*

A STRUCTURAL EQUATION MODELING APPROACH TO UNDERSTANDING THE CARDIOVASCULAR EFFECTS OF ANTIRETROVIRAL THERAPY (ART) INITIATION: RESULTS FROM A PHASE III CLINICAL TRIAL

Carlee B. Moser*, Harvard School of Public Health

Judith S. Currier, University of California, Los Angeles

James H. Stein, University of Wisconsin School of Medicine and Public Health

Howard N. Hodis, University of Southern California

Michael P. Dube, University of Southern California

Todd T. Brown, Johns Hopkins University School of Medicine

Grace A. McComsey, Case Western Reserve University School of Medicine

ACTG A5260s, a prospective-metabolic-substudy of A5257, was designed to evaluate the cardiovascular effects of antiretroviral therapy (ART) initiation with contemporary ART regimens. A total of 328 HIV-infected, ART-naive persons at least 18 years of age, without known cardiovascular disease were randomly assigned to tenofovir disoproxil fumarate-emtricitabine (TDF/FTC) plus either atazanavir-ritonavir(ATV/r), darunavir/ritonavir(DRV/r), or raltegravir(RAL). Annual assessments of follow-up included carotid artery intima-media thickness (CIMT) and several inflammatory and immunological biomarkers. Contrary to the primary hypothesis the study results demonstrated a slower rate of CIMT progression with ATV/r compared to DRV/r; the rate with RAL was intermediate. It was hypothesized that treatment-induced early changes in biomarkers mediate these differences in longer-term CIMT progression. However, traditional modeling approaches are problematic due to the large number of biomarkers and temporal relationships of treatment, biomarkers, and outcomes. To examine this hypothesis, exploratory factor analysis was used to group biomarkers by latent factors, and examine whether these latent factors mediate

treatment-associated changes in outcomes.This analysis considered 22 biomarkers at an early time-point (either 24 or 48 weeks on-treatment) to create latent factors, and structural equation models to relate treatment, baseline CD4 cell count, baseline viral load, and the latent factors with longitudinal CIMT progression.

email: *cmoser@sdac.harvard.edu*

58. ADVANCES AND CHALLENGES IN BIOMARKER STUDIES

A PARADIGM FOR CENTER EFFECTS IN BIOMARKER STUDIES

Kathleen F. Kerr*, University of Washington

Allison Meisner, University of Washington

When building predictive models, most variables are candidate predictors regardless of their causal relationship with the outcome of interest. However, in a multi-center study, study center may be predictive of the outcome, yet center is not a candidate predictor because the goal is to develop a predictive model that can be applied to patients from new centers. A common choice in this situation is to ignore center. However, ignoring center can bias the development of the predictive model if there are systematic differences in biomarker measurements across centers, which may be the case for new biomarkers. We introduce a paradigm for considering center effects in predictive modeling contexts that is similar to the taxonomy of confounding variables and precision variables in etiologic studies. We present methods for seeking combinations of biomarkers in the presence of possible center effects, including methods for estimating center-adjusted optimism-corrected estimates of model performance.

email: *katiek@uw.edu*

AN EFFICIENT PROCEDURE TO COMBINE BIOMARKERS WITH LIMITS OF DETECTION FOR RISK PREDICTION

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

Diego Tomassi, Instituto de Matemática Aplicada del Litoral, Argentina

Efstathia Bura, George Washington University

Liliana Forzani, Universidad Nacional del Litoral,Argentina

Only a few procedures have been proposed so far that address how to combine information from multiple correlated markers that are also left and/or right censored due to lower or upper limits of detection. We extend dimension reduction approaches, specifically likelihood-based sufficient dimension reduction (LDR) to regression or classification with censored predictors. These methods apply generally to any type of outcome, including continuous and categorical outcomes. Using an expectation maximization (EM) algorithm, we find linear combinations that contain all the information contained in correlated markers for modeling and prediction of an outcome variable, while accounting for left and right censoring due to detection limits. We also allow for selection of important variables through penalization. We assess the performance of our methods extensively in simulations and apply them to data from a study conducted to assess associations of 47 inflammatory markers and lung cancer risk and build prediction models.

email: *pfeiffer@mail.nih.gov*

VALIDATION OF RECLASSIFICATION MEASURES - THE ROLE OF CALIBRATION

Nancy R. Cook*, Brigham and Women's Hospital

Risk reclassification methods have become popular in the medical literature as a means of comparing risk prediction models, particularly when adding new biomarkers to established risk factors. It is well-known that assessing measures of improvement in model development data sets may lead to optimistic estimates. The extent of such bias will be described under various model assumptions in training and test sets. The role of model calibration on each of the measures, including the NRI, IDI, and reclassification calibration test, will be examined, as well as the effects of shrinking model coefficients. This will be described for both nested and non-nested models.

email: *ncook@rics.bwh.harvard.edu*

CORRECTING FOR OVER-OPTIMISM IN METRICS OF PROGNOSTIC MODEL IMPROVEMENT

Megan L. Neely*, Duke University and Duke Clinical Research Institute

Michael J. Pencina, Duke University and Duke Clinical Research Institute

Metrics of incremental improvement are often used to compare the performance of a known risk model to an updated version. In practice, researchers often report the apparent performance of a model where the performance metric is estimated using the same data used to estimate the model parameters. It has been shown that this approach can lead to overly optimistic estimates for metrics of individual model performance; however, the behavior of metrics of incremental performance has not been well studied. In this work, we aim to understand the behavior of incremental metrics in this setting when comparing two nested risk models for binary outcomes, with the primary goal of quantifying the level of over-optimism and then evaluating an approach for obtaining point and interval estimates that correct for the positive bias. We focus on the influence of sample size, event rate, and effect size of the novel predictors in the expanded model, including both null and non-null effect sizes. Our proposed ap-

proach is an extension of an approach commonly used for individual performance metrics that not only provides unbiased estimates of incremental improvement, but also provides estimates of precision so that confidence intervals can be constructed - a feature lacking in the currently available approaches.

email: *megan.neely@duke.edu*

59. FUNCTIONAL REGRESSION METHODS AND PERSONALIZED MEDICINE

FUNCTIONAL REGRESSION METHODS FOR DENSELY-SAMPLED BIOMARKERS IN THE ICU

Ciprian Crainiceanu*, Johns Hopkins University

I introduce a class of methods for modeling longitudinal predictors by treating them as functional covariates in regression models. First, I introduce Variable-Domain Functional Regression, which extends the generalized functional linear model by allowing for functional covariates that have subject-specific domain widths. I then propose a blueprint for the inclusion of baseline functional predictors in Cox proportional hazards models. Finally, I propose the Historical Cox Model, which introduces a new way of modeling time-varying covariates in survival models by including them as historical functional terms. Methods were motivated by and applied to a study of association between daily measures of the Intensive Care Unit (ICU) Sequential Organ Failure Assessment (SOFA) score and mortality, and are generally applicable to a large number of new studies that record continuous variables over time.

email: *ccraini1@jhu.edu*

FUNCTIONAL FEATURE CONSTRUCTION FOR PERSONALIZED TREATMENT REGIMES

Eric B. Laber*, North Carolina State University

Robert Pehlman, North Carolina State University

Ana-Maria Staicu, North Carolina State University

Evidence-based personalized medicine formalizes treatment selection as an individualized treatment regime that maps up-to-date patient information into the space of possible treatments. Available patient information may include static features such race, gender, family history, genetic and genomic information, as well as longitudinal information including the emergence of comorbidities, waxing and waning of symptoms, side-effect burden, and adherence. Dynamic information measured at multiple time points before treatment assignment should be included as input to the treatment regime. However, subject longitudinal measurements are typically sparse, irregularly spaced, noisy, and vary in number across subjects. Existing estimators for treatment regimes require equal information be measured on each subject and thus standard practice is to summarize longitudinal subject information into a scalar, ad hoc summary during data pre-processing. This reduction of the longitudinal information to a scalar feature precedes estimation of a treatment regime and is therefore not informed by subject outcomes, treatments, or covariates. Furthermore, we show that this reduction requires more stringent causal assumptions for consistent estimation than are necessary. We propose a data-driven method for constructing maximally prescriptive yet interpretable features that can be used with standard methods for estimating optimal treatment regimes. In our proposed framework, we treat the subject longitudinal information as a realization of a stochastic process observed with error at discrete time points. Functionals of this latent process are then combined with outcome models to estimate an optimal treatment regime. The proposed methodology requires weaker causal assumptions than Q-learning with an ad hoc scalar summary and is consistent for the optimal treatment regime.

email: *laber@stat.ncsu.edu*

ESTIMATION OF OPTIMAL TREATMENT POLICIES AND MARGINAL SCREENING OF FUNCTIONAL PREDICTORS

Ian W. McKeague*, Columbia University

McKeague and Qian (2014) recently introduced a functional regression approach to the problem of estimating optimal personalized treatment policies. Here we discuss an enhancement of this approach involving marginal screening. The idea is to incorporate an adaptive resampling test to screen out components of the functional predictor that have no significant interaction with the treatment. This is joint work with Min Qian.

email: *imckeague@hotmail.com*

DEVELOPING BIOMARKERS FOR BRAIN LESION TRAJECTORIES IN LONGITUDINAL MRI

Elizabeth M. Sweeney, Johns Hopkins University

Russell T. Shinohara*, University of Pennsylvania

Blake E. Dewey, National Institute of Neurological Disorders and Stroke, National Institutes of Health

Matthew K. Schindler, National Institute of Neurological Disorders and Stroke, National Institutes of Health

John Muschelli, Johns Hopkins University

Daniel S. Reich, National Institute of Neurological Disorders and Stroke, National Institutes of Health

Ciprian M. Crainiceanu, Johns Hopkins University

Ani Eloyan, Brown University

Multiple sclerosis (MS) is an immune-mediated disease of the central nervous system in which inflammatory and demyelinating lesions form in the white matter of the brain and spinal cord. Although the accumulation of these lesions throughout the disease course is known to be associated with morbidity and disability, the association between the volume of lesion visible on magnetic resonance imaging (MRI) and clinical outcomes is weak. Despite this, MRI outcomes are standard for clinical trials of new disease-modifying therapies for MS. In this talk, we describe a statistical framework based on functional data analysis for studying longitudinal patterns of lesion development. This new paradigm uses statistical and classical quantitative MRI to assess lesion severity and the immune system's capacity to repair these lesions through the disease course, and provides sensitive biomarkers for measuring treatment effects.

email: *rshi@upenn.edu*

60. BAYESIAN METHODS FOR LARGE-SCALE NON-GAUSSIAN DATA

BAYESIAN MODELING OF HUGE TABLES AND DISCRETE DATA

David B. Dunson*, Duke University

Although there has been abundant consideration of methods for dimensionality reduction and modeling of high-dimensional Gaussian data, and more broadly real-valued continuous vectors, relatively little consideration has been given to high-dimensional tables and discrete data. I focus on problems arising in analysis of massive scale categorical and count data, proposing recent scalable Bayesian solutions relying on low rank factorizations and novel classes of shrinkage priors. Some basic theory is provided, efficient algorithms are outlined, and I consider several interesting motivating biomedical examples including to epidemiology and genome wide sequencing.

email: *dunson@duke.edu*

BAYESIAN MODELS OF HIGH-DIMENSIONAL COUNT DATA

Marina Vannucci*, Rice University

Michele Guindani, University of Texas MD Anderson Cancer Center

Many of the real applications prevalent in the modern data science involve heterogeneous and mixed data (e.g. count, binary, continuous, skewed continuous, among other data types). In this talk we will consider hierarchical Bayesian models for high-dimensional count data that incorporate variable selection. Zero-inflation, skewness, and overdispersion all cause difficulties when modeling count data. We will first look at the problem of clustering a high-dimensional matrix of count data and develop a Bayesian nonparametric hierarchical Poisson mixture model that accounts for the overdispersion observed across samples as well as across multiple features. The model formulation incorporates a feature section mechanism and prior distributions that appropriately account for identifiability constraints on the model parameters. If time allows, we will also explore extension of the methodologies to Poisson and Dirichlet-Multinomial regression models. The development of the methods will be motivated by applications in different fields, including bag-of-words examples from machine learning and high-throughput datasets from integrative genomics.

email: *marina@rice.edu*

VALID STATISTICAL ANALYSES AND REPRODUCIBLE SCIENCE IN THE ERA OF HIGH-THROUGHPUT

Edoardo M. Airoldi*, Harvard University

High-throughput technology (eg, sequencing, mass spec allows us to quantify biological mechanisms at a resolution that array technology and small scale experiments cannot. In the next 5-10 years, a substantial portion of biological research is expected to leverage some of these technologies. This flexibility comes with a price, however. Modern high-throughput instrumentation relies on built-in data collection protocols that are often biased. (For instance, a mass spec selects the most abundant ions, at an early stage of the measurement process, for further analysis.) The major unexpected consequence of such protocols is that they carry information about those quantities we are interested in estimating (absolute protein abundance, in the mass spec example). Scientists that do not account for this information during analysis, whether by counting or estimation using a statistical model, will likely base their scientific conclusions on misleading numbers, even in simple experimental conditions. This statistical issue is poorly understood by practitioners and amateur statisticians alike. It is arguably the main challenge we need to tackle to produce valid scientific conclusions in the era of high-throughput technology. I'll provide two illustrations in mass spectrometry and genomics.

email: *airoldi@fas.harvard.edu*

61. CANCER APPLICATIONS

P53-BASED STRATEGY TO REDUCE HEMATOLOGICAL TOXICITY OF CHEMOTHERAPY: A PILOT STUDY

Chul S. Ha, University of Texas Health Science Center, San Antonio

Joel Michalek*, University of Texas Health Science Center, San Antonio

Richard Elledge, University of Texas Health Science Center, San Antonio

Kevin R. Kelley, University of Texas Health Science Center, San Antonio

Suthakar Ganapathy, University of Texas Health Science Center, San Antonio

Su Hang, University of Texas Health Science Center, San Antonio

Carol A. Jenkins, University of Texas Health Science Center, San Antonio

Athanassios Argiris, University of Texas Health Science Center, San Antonio

Ronan Swords, University of Texas Health Science Center, San Antonio

Tony Y. Eng, University of Texas Health Science Center, San Antonio

P53 activation is the primary mechanism underlying pathological responses to DNA-damaging agents such as chemotherapy and radiotherapy. Study objectives were to: 1) define the lowest safe dose of arsenic trioxide that blocks p53 activation in patients and 2) assess the potential of LDA to decrease hematological toxicity from chemotherapy. Patients scheduled to receive a minimum of 4 cycles of myelosuppressive chemotherapy were eligible. For objective 1, dose escalation of LDA started at 0.005mg/kg/day for 3 days. This dose satisfied objective 1 and was administered before chemotherapy cycles 2, 4 and 6 for objective 2. CBC was compared between the cycles with and without LDA pretreatment. Subjects received arsenic at cycles 2, 4 and 6 and no arsenic at cycles 1, 3, and 5. Of a total of 30 evaluable patients, 26 were treated with 3-week cycle regimens and form the base of our analyses. The mean white blood cell, hemoglobin and absolute neutrophil counts were significantly higher in the suppressed group relative to the activated group. These data support the proof of concept that suppression of p53 could lead to protection of normal tissue and bone marrow in patients receiving chemotherapy.

email: *michalekj@uthscsa.edu*

USING IMRE AND ANOVA TO SELECT MicroRNAs FOR PREDICTING PROSTATE CANCER RECURRENCE

Qi Wang*, North Dakota State University

Bin Guo, North Dakota State University

Yarong Yang, North Dakota State University

Imputed microRNA regulation based on weighted ranked expression and putative microRNA targets (IMRE) is a method of predicting microRNA regulation from genome-wide gene expression as well as predicting microRNA putative targets. A false discovery rate for each microRNA is calculated using the expression of the microRNA putative targets to analyze the regulation between different conditions. ANOVA is a statistical method used to analyze the differences between group means. It tests if the means of several groups are equal. We apply the IMER and ANOVA methods on a prostate cancer gene expression dataset with 596 men of three different phenotypes: PSA (Prostate-Specific Antigen recurrence), Systemic (Systemic disease progression) and NED (No Evidence of Disease). A group of microRNAs are selected for experimental tests for the PSA recurrence and Systemic disease progression.

email: *qi.wang.1@ndsu.edu*

IMPLEMENTATION OF A 2-STAGE CROSSOVER CORRECTION IN ANALYSIS OF OVERALL SURVIVAL (OS): AN EXAMPLE IN ONCOLOGY

Ruifeng Xu*, Merck & Co., Inc.

Jingshu Wang, Merck & Co., Inc.

James M. Pellissier, Merck & Co., Inc.

KN002 is a phase 2 RCT of pembrolizumab vs. investigator-choice chemotherapy in ipilimumab-refractory melanoma. In this ongoing study, ITT analyses at the second interim analysis, showed a numerical trend of improved OS in favor of pembrolizumab 2 mg/kg Q3W with the HR of 0.88 (p=0.229) vs. the control. However, the effect was confounded by high rate (49%) of crossover post-progression by the control group crossing to receive pembrolizumab. We detail the application of the 2-stage method (Latimer et al. 2014) for cross-over adjustment to the OS data. For the crossover adjusted model, "time of progression" was selected as a secondary baseline after considering complexities related to protocol mandated post-progression washout time. Covariates were selected and the 2-stage method applied. Crossover corrected HR was 0.63 (p=0.007). Model validation showed adjusted control OS closely followed historical controls. Sensitivity analyses testing the impact of the selected secondary baseline and the modeled covariates showed results to be robust. The 2-stage crossover adjustment method performed well. Rigorous validation of the crossover correction methodology for OS is recommended.

email: *ruifeng_xu@merck.com*

MODELING MULTIPLE PRIMARY CANCERS OVER TIME USING NON-HOMOGENEOUS POISSON PROCESS

Jialu Li*, University of Texas MD Anderson Cancer Center

Seung Jun Shin, Korea University

Wenyi Wang, University of Texas MD Anderson Cancer Center

A common phenomenon in cancer is that for the same individual, multiple sites may present primary cancer at different times in life. In Li-Fraumeni syndrome(LFS), a rare pediatric cancer syndrome, TP53 mutation carriers are known to have a high probability of developing second primary cancer than the non-carriers. Modeling the development of multiple primary cancers is therefore desired for better clinical management of LFS. To this end, we have devel-

oped a non-homogeneous Poisson process model with multiplicative rate function in order to account for primary cancer events occurred over time. We constructed a novel familywise likelihood under the Poisson process, which allows for inclusion of all family history information even when the individual genotype is missing. We used Markov chain Monte Carlo algorithm to estimate model parameters, and derived age-at-onset penetrance for single and multiple primary cancers given mutation status, respectively. We applied our method to a pediatric sarcoma cohort collected at MD Anderson Cancer Center from 1944 to 1982. Our penetrance estimates are based on a total of 3,686 individuals from 189 families, and are consistent with SEER estimates and previous studies on LFS.

email: *jialu.lilee@gmail.com*

ASSESSING INTRA-TUMOR HETEROGENEITY AND TRACKING LONGITUDINAL AND SPATIAL CLONAL EVOLUTION BY NEXT-GENERATION SEQUENCING

Yuchao Jiang*, University of Pennsylvania

Andy J. Minn, University of Pennsylvania

Nancy R. Zhang, University of Pennsylvania

Cancer is a disease driven by genetic and epigenetic alterations that follows Darwinian evolution. Recently, there have been increasing efforts to sequence the tumor from the same patient at multiple time points and/or from multiple spatially separated resections. Different snapshots of the same tumor have proved invaluable for inferring the tumor's clonal history. We propose a method, Canopy, for reconstructing subclonal phylogeny utilizing both copy number alterations (CNAs) and single nucleotide alterations (SNAs) from one or more samples derived from a single patient. Canopy provides a general mathematical framework that enumerates all possible CNA-SNA phases and gives confidence assessments of all possible phylogenetic configurations. On a whole-exome study of a transplantable metastasis model derived from cell line MDA-MB-231, Canopy successfully deconvolutes the mixed-cell sublines, using the single-cell sublines as ground truth. On a whole-genome sequencing dataset of the breast cancer tumor and its subsequent metastatic xenograft, Canopy's inferred clonal phylogeny is confirmed by single-cell sequencing. Finally, through simulations, we explore the effects of various parameters on deconvolution accuracy, and evaluate performance with comparison against existing methods. Collectively, Canopy provides a rigorous foundation for statistical inference on repeated sequencing experiments from evolving populations delineated temporally and spatially.

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

PATHWAY-BASED DIFFERENTIAL NETWORK ANALYSIS IN CANCER

Min Jin Ha*, University of Texas MD Anderson Cancer Center

Veerabhadran Baladandayuthapani, University of Texas MD Anderson Cancer Center

Kim-Anh Do, University of Texas MD Anderson Cancer Center

Cancer progression and development are initiated by aberrations in various molecular networks through coordinated changes across multiple genes and pathways. It is important to understand how these networks change under different stress conditions and/or patient-specific groups to infer differential patterns of activation and inhibition. Existing methods are limited to correlation networks that are independently estimated from separate group- specific data and without due consideration of relationships that are conserved across multiple groups. We propose a pathway-based differential network analysis in genomics (DINGO) model for estimating group-specific networks as well as making inference on the differential networks. DINGO jointly estimates the group-specific conditional dependencies by decomposing them into global and group-specific components. The delineation of these components allows for a more refined picture of the major driver and passenger events in the elucidation of cancer progression and development. Simulation studies demonstrate that DINGO provides more accurate group-specific conditional dependencies than achieved by using separate estimation approaches. We apply DINGO to key signaling pathways in glioblastoma to build differential networks for long-term survivors and short-term survivors in The Cancer Genome Atlas (TCGA). The hub genes found by mRNA expression, DNA copy number, methylation and microRNA expression, reveal several important roles in glioblastoma progression.

email: *mjha@mdanderson.org*

CELL TYPE-SPECIFIC DECONVOLUTION OF HETEROGENEOUS TUMOR SAMPLES WITH IMMUNE INFILTRATION USING EXPRESSION DATA

Zeya Wang*, Rice University

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

Jaeil Ahn, Georgetown University

Bo Li, Harvard University

Wei Lu, University of Texas MD Anderson Cancer Center

Ximing Tang, University of Texas MD Anderson Cancer Center

Ignacio I. Wistuba, University of Texas MD Anderson Cancer Center

Chris C. Holmes, University of Oxford

Wenyi Wang, University of Texas MD Anderson Cancer Center

Tumor tissue samples comprise of a mixture of cancerous and surrounding stromal cells. Understanding tumor heterogeneity is crucial to analyzing gene signatures associated with cancer prognosis and treatment decisions. Numerous computational approaches previously developed all have their limitations to deconvolute heterogeneous tumor samples. We have significantly developed a three-component deconvolution model, DeMix-T, that can explicitly account for a third component such as the immune cell compartment and is able to address this challenging problem when the observed signals are assumed to come from a mixture of three cell compartments, infiltrating immune cells, the tumor microenvironment and cancerous tis- sues, instead of two. DeMix-T is computationally feasible when it is needed to compute high-dimensional integrals and involves a novel two-stage filtering method that yields accurate estimates of cell purities and compartment-specific expression profiles. Simulations and real data analyses have demonstrated the good performance of our method. Compared with other deconvolution tools, DeMix-T can be applied more widely and provides deeper insight into cancer biomarker studies. It allows for a further understanding of immune infiltration in cancer and assists in the development of novel prognostic markers and therapeutic strategies.

email: *zw17@rice.edu*

62. HETEROGENEOUS TREATMENT EFFECTS

ON CLINICAL TRIALS WITH A HIGH PLACEBO RESPONSE RATE

George Chi, Janssen Research & Development, LLC

Pilar Lim*, Janssen Research & Development, LLC

The basic reason for the failure of many standard randomized parallel placebo-controlled clinical trials with high placebo response rate is that the observed relative treatment difference only provides an estimate of the apparent treatment effect since the true treatment effect has been diminished by the presence of a substantial proportion of placebo responders in the population. Analogous to an active control trial, the true treatment effect cannot be measured by the relative treatment difference. An appropriate assessment of the true treatment effect is critical for making a risk/benefit analysis and dosage recommendation. The primary purpose of this talk is to propose a method for adjusting the apparent treatment effect to account for the high placebo response rate within the framework of a doubly randomized delayed start design.

email: *plim@its.jnj.com*

USING IMRE AND DUAL KS TO SELECT MicroRNAs FOR PREDICTING PROSTATE CANCER RECURRENCE

Yarong Yang*, North Dakota State University

Qi Wang, North Dakota State University

Imputed microRNA regulation based on weighted ranked expression and putative microRNA targets (IMRE) is a method of predicting microRNA regulation from genome-wide gene expression as well as predicting microRNA putative targets. A false discovery rate for each microRNA is calculated using the expression of the microRNA putative targets to analyze the regulation between different conditions. Dual KS is an efficient analytic methodology that can identify class specific highly parsimonious gene signatures. We apply the IMER and Dual KS methods on a prostate cancer gene expression dataset with 596 men of three different phenotypes: PSA (Prostate-Specific Antigen recurrence), Systemic (Systemic disease progression) and NED (No Evidence of Disease). A group of microRNAs are selected for experimental tests for the PSA recurrence and Systemic disease progression.

email: *yarong.yang@ndsu.edu*

ESTIMATING TREATMENT EFFECT IN TIME TO RELAPSE WHEN PATIENTS SWITCH TREATMENT

Miao Lu*, University of Virginia

Jian Han, Genentech, Inc.

Randomized controlled trials (RCTs) are widely used to evaluate the effects of a new treatment over a control treatment. In reality, it's common for patients who were randomized to the control group, switch onto the experimental treatment at some point during follow up. Treatment switches may occur for a number of reasons, which are ethical and practical, and related to patient's prognosis. The question of interest is what would have been the survival treatment effect had no patients in the control group switched. This paper reviews available methods about treatment switch, and applies iterative parameter estimation (IPE) method in the Multiple Sclerosis treatment analysis. Moreover, it tests the performance of the method by extensive simulations. Additionally, we extend the existing method to incorporate baseline covariates adjustment, and relatively high percent of censoring. Last but not least, similar idea can be used in different type of switching like patients from treatment group switched to control group.

email: *ml4ey@virginia.edu*

INFERENCE ON SUBGROUPS AND ALL-COMERS COGNIZANT OF LOGICAL RELATIONSHIPS AMONG EFFICACY PARAMETERS

Szu-Yu Tang*, Ventana Medical Systems, Inc. (Roche Group)

Yi Liu, Millennium: The Takeda Oncology Company

Jason Hsu, Eli Lilly & Company and The Ohio State University

In one aspect of personalized medicine development, the patient population is thought of as a mixture of two subgroups that might derive differential efficacy, given treatment versus control. An important decision to make is whether to target the entire patient population (so-called all-comers), or just a subgroup of the patients. There are logical relationships among efficacy parameters in the subgroups and all-comers. This presentation shows the Partition Principle in multiple testing can formulate null hypotheses that respect such logic, and discusses to what extent statistical inference should respect these logical relationships.

e-mail: *tang.142@buckeyemail.osu.edu*

LOGICAL INFERENCE ON TREATMENT EFFICACY IN SUBGROUPS AND THEIR MIXTURES

Ying Ding*, University of Pittsburgh

Hui-Min Lin, Takeda Pharmaceuticals

Jason C. Hsu, Eli Lilly & Company and The Ohio State University

Measuring treatment efficacy in a mixture population is a fundamental problem in tailored drug development, in deciding which subgroup or combination of subgroups to treat. Such a development process typically involves comparing a new drug with a control through randomized clinical trials, and treatment efficacy is the relative effect between the new drug and the control. A fundamental consideration in this inference process is that the logical relationships between treatment efficacy in subgroups and their combinations should be respected. We show that some commonly used efficacy measures are not suitable for a mixture population. We also show that, while it is important to adjust for imbalance in the data using least squares means (LSmeans) (not marginal means) estimation, current practice over-extends the LSmeans concept when estimating the efficacy in a mixture population. Proposing a subgroup mixable estimation principle, we develop a simultaneous inference procedure, with appropriate efficacy measures, to confidently infer efficacy in subgroups and their mixtures.

e-mail: *yingding@pitt.edu*

CONFIDENT EFFECT OF A SNP ON THE EFFICACY OF A DRUG

Jason C. Hsu*, Eli Lilly & Company and The Ohio State University

Ying Ding, University of Pittsburgh

Ying Grace Li, Eli Lilly & Company

Stephen J. Ruberg, Eli Lilly & Company

In testing for SNPs predictive of treatment efficacy, a common practice is as follows. For each SNP, several tests (including those for dominant, recessive, and additive effects) are executed. The minimum p-value of these tests is taken to represent the potential significance of that SNP. The SNPs are then ranked according to their p-values, from smallest to largest. Those with p-values smaller than a threshold meant to control a multiple testing error rate such as the False Discovery Rate (FDR) or per family error rate are then inferred to be "significant". We suggest an alternative strategy that is more informative for the purpose of drug development. Set a confidence level adjusted for the multiplicity of SNPs according to the error rate of choice (e.g., FDR, or per family). For each SNP, compute simultaneous confidence intervals at that level for dominant, recessive, and additive effects on the clinical response. In Type II diabetes, for instance, they would be confidence intervals for mean difference between treatment and control of HbA1c reduction from baseline. Define the "confident effect" of a confidence interval (CI) as the minimum distance of points in that interval from zero (so the confident effect of a CI containing zero is zero). The maximum confident effect is taken to represent the potential significance of that SNP. Report and plot the maximum confident effects of the SNPs, ordered from largest to smallest. This ranking is different from p-value ranking, and is informative of both the nature (dominant, recessive, additive) and the size of the effect.

e-mail: *jch@stat.osu.edu*

A PREDICTIVE ENRICHMENT PROCEDURE TO IDENTIFY POTENTIAL RESPONDERS TO A NEW THERAPY FOR RANDOMIZED, COMPARATIVE CONTROLLED CLINICAL STUDIES

Junlong Li, Harvard University

Lihui Zhao*, Northwestern University

L**u Tian**, Stanford University

Tianxi Cai, Harvard University

Brian Claggett, Brigham and Women's Hospital

Andrea Callegaro, GlaxoSmithKline Vaccines

Benjamin Dizier, GlaxoSmithKline Vaccines

Bart Spiessens, GlaxoSmithKline Vaccines

Fernando Ulloa-Montoya, GlaxoSmithKline Vaccines

L. J. Wei, Harvard University

To evaluate a new therapy versus a control via a randomized trial, due to heterogeneity of the study patient population, a pre-specified, predictive enrichment procedure may be implemented to identify an "enrichable" subpopulation. For patients in this subpopulation, the therapy is expected to have a desirable overall risk-benefit profile. To develop and validate such a "therapy-diagnostic co-development" strategy, a three-step procedure may be conducted with three independent data sets from a series of similar studies or a single trial. At the first stage, we create various candidate scoring systems based on the baseline characteristics via, for example, parametric models using the first data set. Each individual score reflects an anticipated

treatment difference for future patients who share similar baseline profiles. At the second step, a potentially enrichable subgroup is identified using the totality of evidence from these scoring systems. At the final stage, we validate such a selection via two-sample inference procedures for assessing the treatment effectiveness statistically and clinically with the third data set, the so-called holdout sample. When the study size is not large, one may combine the first two steps using a "cross-training-evaluation" process. The entire enrichment procedure is illustrated with the data from a cardiovascular trial.

email: *lihui.zhao@northwestern.edu*

63. HIGH DIMENSIONAL DATA APPLICATIONS

LINEAR SHRINKAGE REVISITED: POSITIVE-DEFINITE MODIFICATION OF LARGE-DIMENSIONAL COVARIANCE MATRIX ESTIMATORS WITH APPLICATIONS TO REHABILITATIVE SPEECH TREATMENT OF PATIENTS WITH PARKINSON'S DISEASE

Young-Geun Choi*, Seoul National University

Johan Lim, Seoul National University

Most patients with Parkinson's disease experience vocal performance degradation and periodically undergo personalized rehabilitation therapies with speech experts. In an attempt to establish an automated therapeutic algorithm for those patients' speech, Tsanas et al. (2014) collected data on patients' speech and experts' evaluation categorized into "acceptable" and "non-acceptable". This data calls for large-dimensional classification, typical when developing diagnostic or therapeutic algorithms in digital healthcare. One of the relevant methodologies is linear minimax probability machine (LMPM) (Lanckriet et al, 2002). A challenge in implementing LMPM for large-dimensional data is the lack of positive-definiteness (PDness) of the regularized covariance matrix estimators. Existing solutions incorporated steps guaranteeing PDness in the optimization process, computationally demanding. We propose a two-stage approach called linear shrinkage for positive-definiteness (LSPD), where covariance matrix regularization and its conversion to PDness are two separate steps. We show that the resulting estimator preserves the asymptotic properties of the initial estimator if the shrinkage parameters are carefully selected. We conducted simulation studies to evaluate the finite-sample and computational properties of the proposed estimator and existing estimators. Finally, we applied the proposed method to rehabilitative speech treatment of patients with Parkinson's disease and demonstrated substantial improvement in the LMPM classification accuracy.

email: *eumjangi@snu.ac.kr*

MULTIVARIATE TEST FOR HIGH DIMENSIONAL COMPOSITIONAL CHANGES IN PAIRED MICROBIOME STUDIES

Ni Zhao*, Fred Hutchinson Cancer Research Center

Xiang Zhan, Fred Hutchinson Cancer Research Center

Michael Wu, Fred Hutchinson Cancer Research Center

Human microbiome composition is subject to change in response to events such as disease, antibiotic treatment, stress, injury, and changes in diet. However, statistical testing for the global change in paired compositional data are challenging due to 1) the microbiome compositional data provides information only about the relative abundances of different bacterial species, which can be correlated in very complex way; 2) the number of microbiome species are large compared to the sample size. In this paper, we propose a multivariate test procedure to test for high dimensional compositional changes in paired microbiome studies using a regularized Hotelling's T2statistics. We compare our proposed method to a few ad-hoc approaches for paired compositional data and show improved power and better control of type I error.

email: *nzhao@fhcrc.org*

KERNEL-BASED NONPARAMETRIC TESTING IN HIGH-DIMENSIONAL DATA WITH APPLICATIONS TO GENE SET ANALYSIS

Tao He*, San Francisco State University

Ping-Shou Zhong, Michigan State University

Yuehua Cui, Michigan State University

Vidyadhar Mandrekar, Michigan State University

This paper considers testing a nonparametric function of high-dimensional variates in a reproducing kernel Hilbert space, which is a function space generated by a positive definite kernel function. We propose a test statistic to test the nonparametric function under the high-dimensional setting. The asymptotic distributions of the test statistic are derived under the null hypothesis and a series of local alternative hypotheses, in the \large p, small n\ setup. Extensive simulation studies and a real data analysis were conducted to evaluate the performance of the proposed method.

email: *hetao@sfsu.edu*

COVARIANCE ENHANCED SCREENING FOR ULTRAHIGH-DIMENSIONAL CLASSIFICATION

Yanming Li*, University of Michigan

Kevin Ke, University of Michigan

Ji Zhu, University of Michigan

Yi Li, University of Michigan

Classification for high-dimensional feature space has been extensively studied in recent years (Fan and Fan, 2008; Guo, 2010; Xu et al., 2014), there are, however, still challenges remain. First, most current high-dimensional classifiers can not be applied to ultrahigh-dimensional cases, where the number of features is of at least exponential order of the sample size and which are often the cases in modern cancer genomic studies. Secondly, many high-dimensional classifiers assume the independence rule and ignore the inter-feature correlation and therefore are not able to detect signals that are marginally not discriminative but jointly informative. Here we propose a multivariate screening method for ultrahigh-dimensional classification for weak and sparse signals by incorporating the inter-feature correlation. Under some mild regularity conditions, we show that the proposed screening method assumes the sure screening property, well controls the false positives, and achieves an asymptotic minimal misclassification rate. We also show that the proposed method provides a improved phase diagram compared to Jin (2009) in the sense that it is able to identify marginally weak informative signals that would had been labeled as impossible to classify by Jin (2009). The performance of the proposed method is evaluated by extensive simulations and we apply the method to a renal transplant data for post-transplantation renal functional type classification.

email: *liyanmin@umich.edu*

DISSECTING THE GENE-ENVIRONMENT INTERACTIONS: A ROBUST PENALIZATION APPROACH ACCOUNTING FOR HIERARCHICAL STRUCTURES

Cen Wu*, Kansas State University and Yale University

Yu Jiang, University of Memphis

Shuangge Ma, Yale University

Identification of gene-environment (G×E) interactions associated with disease phenotypes has been a great challenge in high throughput cancer studies. Existing marginal identification methods have suffered from not being able to accommodate the joint effects of a large number of genetic variants, while the joint-effects approaches have been limited by using inefficient selection techniques, by failing to respect the "main effect, interaction" hierarchy, and by assuming no data contamination. We propose an efficient penalization approach to identify important G×E interactions and main effects while accounting for hierarchical structures between the two type of effects. Possible data contamination has been taken care of by adopting the least absolute deviation (LAD) loss function. The advantage of the proposed approach over the alternatives has been demonstrated in both simulation study and a case study on a lung cancer prognosis study with gene expression measurements and clinical covariates under the AFT (accelerated failure time) model.

email: *wucen@ksu.edu*

DO-OVER: REPLICATES IN HIGH DIMENSIONS, WITH APPLICATIONS TO LATENT VARIABLE GRAPHICAL MODELS

Kean Ming Tan*, Princeton University

Yang Ning, Princeton University

Daniela Witten, University of Washington

Han Liu, Princeton University

In classical statistics, much thought is put into experimental design and data collection. However, in the high-dimensional setting, experimental design has been less of a focus, and often the data collected are not sufficient or appropriate for answering the scientific question of interest. In this paper, we stress the importance of collecting multiple replicates for each subject in the high-dimensional setting via a case study on learning the structure of a high-dimensional graphical model with latent variables, under the assumption that the latent variables take on a constant value across replicates within each subject. By collecting multiple replicates for each subject, we are able to estimate the conditional dependence relationships among the observed variables given the latent variables. To test the null hypothesis of conditional independence between two observed variables, we propose a pairwise decorrelated score test. Theoretical guarantees are established for parameter estimation and for the pairwise decorrelated score test. Through numerical studies, we investigate the finite sample performance of our proposal relative to existing proposals across different types of graphical models with latent variables. Finally, we apply the proposed method to a brain-imaging dataset.

email: *tan.keanming@gmail.com*

SPARSE LINEAR DISCRIMINANT ANALYSIS IN STRUCTURED COVARIATES SPACE

Sandra Safo*, Emory University

Qi Long, Emory University

Linear discriminant analysis (LDA) is a classical multivariate analysis tool popularly used for many classification problems. LDA has limitations in the high dimensional framework as it is usually of interest to select only a fraction of the variables for improved classification accuracy. Several methods for sparse LDA have been proposed in the literature. Although these methods have proven useful in various applications, their main drawback is failure to account for prior biological knowledge. In this paper, we propose a novel structured sparse LDA method that overcomes this limitation by incorporating biological information in the form of graphical networks. We compare our method to existing sparse LDA approaches via simulation studies and real data analysis using gene expression data from cardiovascular disease and breast cancer studies.

email: *seaddosafo@gmail.com*

64. MACHINE LEARNING

LAGGED KERNEL MACHINE REGRESSION FOR IDENTIFYING TIME WINDOWS OF SUSCEPTIBILITY TO COMPLEX METAL MIXTURES

Shelley H. Liu*, Harvard University
Jennifer F. Bobb, Harvard School of Public Health
Kyu Ha Lee, Harvard School of Public Health
Chris Gennings, Mount Sinai Hospital
Birgit Claus Henn, Boston University School of Public Health
Robert O. Wright, Mount Sinai Hospital
Lourdes Schnaas, Instituto Nacional De Salud Publica, Mexico
Martha Tellez Rojo, Instituto Nacional De Salud Publica, Mexico
Manish Arora, Mount Sinai Hospital
Brent Coull, Harvard School of Public Health

A critical public health concern is the impact of neurotoxic chemicals on children's health; exposures to metal mixtures during early life may impact cognitive function, and there may exist critical time intervals during which vulnerability is increased. However, there is a lack of statistical methods to study time-varying exposures of complex toxicant mixtures. Therefore, we develop a flexible statistical method, Lagged Kernel Machine Regression (LKMR), to identify critical exposure windows of chemical mixtures that accounts for complex non-linear and non-additive effects of the mixture at any given exposure window. LKMR is a Bayesian hierarchical model that estimates how the effects of mixture exposures change with the exposure window using a novel grouped, fused Lasso for Bayesian shrinkage. Simulation studies demonstrate the performance of LKMR under realistic exposure-response scenarios, and demonstrate large gains over approaches that consider each critical window separately, particularly when serial correlation among the time-varying exposures is high. We apply LKMR to analyze associations between neurodevelopment and metal mixtures in PROGRESS, a prospective cohort study on metal mixture exposures and neurodevelopment conducted in Mexico City. Our results indicate that the effect of manganese exposure is dependent on exposure timing, and that these manganese effects interact with lead exposure.

email: *shelleyliu@fas.harvard.edu*

A GROUP-SPECIFIC RECOMMENDER SYSTEM

Xuan Bi*, University of Illinois, Urbana-Champaign
Annie Qu, University of Illinois, Urbana-Champaign
Junhui Wang, City University of Hong Kong
Xiaotong Shen, University of Minnesota

Recommender systems have many applications such as in marketing industry, travel, entertainment, online reviews and shopping. In this paper, we propose a group-specific method to utilize cluster information from users and items that share similar missing patterns under the singular value decomposition framework. The new approach is effective for the "cold-start" problem, where, in the testing set, majority responses are obtained from new users or for new items, and their preference information is not available from the training set. In addition, since this type of data involves large-scale customer records, traditional algorithms are not computationally scalable. To implement the proposed method, we propose a new algorithm that embeds a back-fitting algorithm into alternating least squares, which avoids large matrices operation and big memory storage, and therefore makes it feasible to achieve scalable computing. Our simulation studies and MovieLens data analysis all indicate that the proposed group-specific method improves prediction accuracy quite significantly compared to existing competitive methods and algorithms.

email: *xuanbi2@illinois.edu*

A GENERAL UNIMODAL NULL DISTRIBUTION WITH APPLICATIONS TO CLUSTER SIGNIFICANCE TESTING

Erika Helgeson, University of North Carolina, Chapel Hill
Eric Bair*, University of North Carolina, Chapel Hill

In many applications of interest, one seeks to discover if any homogeneous subgroups (i.e., clusters) are present in the data. Many clustering methods exist, but in general it is difficult to determine if the clusters identified by these methods represent truly distinct subgroups in the data or are merely a spurious finding. Testing the null hypothesis that no clusters exist in the data requires the choice of an appropriate unimodal null distribution for the data. We propose a novel null distribution for this problem using kernel density estimation that does not require the data to follow any particular distribution. The significance of a putative set of clusters can be evaluated by comparing the within-cluster sum of squares of the original data to that produced by clustering under this null distribution. We find that our method can accurately test for the presence of clustering quickly and accurately even when the number of features is high.

email: *ebair@email.unc.edu*

RANDOM FORESTS: HOW A CHANCE DRIVEN LEARNING MACHINE DOES SO SPECTACULARLY WELL

Dan Steinberg*, Salford Systems
Adele Cutler, Utah State University

RF is a next generation learning machine based on partially or totally randomly generated decision trees. Each individual tree is a poor predictor of the target but the ensemble of a large number of these trees has proven good enough to win Kaggle competitions and solve many real world problems. This talk presents an overview of the core ideas behind the Random Forest, illustrates its predictive power on a competition data set, explains why the technology works, and discusses the strengths and weaknesses, and types of problems for which RF is best suited. 1. What is a Random Forest? How randomness is incorporated into the learning process by repeatedly training on different rows of data and by considering different subsets of features at each decision node in a tree. Bootstrap samples, OOB: Records included and records excluded from the training of a specific tree ("Out of Bag"), Selecting predictive features at random and how we vary the degree randomness from none to total. 2. What kinds of problems can RF be used to solve. Classification, Regression, Clustering, Outlier and Anomaly Detection. 3. How to set up an RF model: what controls really matter. Number of features considered at each decision node, size of the training sample extracted from the master database, tree size limits. Adapting RF to BigData via very small sample extracts. 4. Binary classification example: predicting credit risk — who repays their loan and who does not. What we learn from RF that we would not learn from other learning machines. 5. Why RF works. The wisdom of crowds applied to trees. Trees, unlike mathematically formulated models, are nothing more than selective descriptions of data. Each tree offers a differently cast description of the data and incorporates a form of nearest neighbor classifier. With sufficient RF trees the average predictive accuracy can become better than that typically delivered by a professional statistician. 6. Parallel RF. Easy ways to get RF models computed rapidly.

email: *lisas@salford-systems.com*

ROBUST LEARNING FOR OPTIMAL TREATMENT DECISION WITH NP-DIMENSIONALITY

Chengchun Shi*, North Carolina State University

Rui Song, North Carolina State University

Wenbin Lu, North Carolina State University

In order to identify important variables that are involved in making optimal treatment decision, Lu et al. (2013) proposed a penalized least squared regression framework for a fixed number of predictors, which is robust against the misspecification of the conditional mean model. Two problems arise: (i) in a world of explosively big data, effective methods are needed to handle ultra-high dimensional dataset, for example, with the dimension of predictors is of the non-polynomial (NP) order of the sample size; (ii) both the propensity score and conditional mean models need to be estimated from data under NP dimensionality. In this paper, we propose a two-step estimation procedure for deriving the optimal treatment regime under NP dimensionality. In both steps, penalized regressions are employed with folded-concave penalty function, where the conditional mean model of the response given predictors may be misspecified. The asymptotic properties, such as weak oracle properties, selection consistency and oracle distributions of the estimators are investigated. In addition, we study the limiting distribution of the estimated value function for the obtained optimal regime. Empirical performance of the proposed method is evaluated by simulations and an application to a depression dataset from the STAR*D study.

email: *cshi4@ncsu.edu*

REGION BASED MEDIATION TEST OF DNA METHYLATION USING KERNEL MACHINE REGRESSION

Jincheng Shen*, Harvard School of Public Health

Xihong Lin, Harvard School of Public Health

Mediation analysis provides a powerful tool in identifying the missing pieces of a given causal relationship. With development of genetic and genomic technologies, it is of increasing interest in biomedical studies to explore the genetic and epigenetic mechanisms that mediate a disease causing process. Epigenetic studies for complex diseases have shown that the effect of many risk factors are likely to be mediated in a collaborative fashion through multiple probes in a genomic region, such as CpG islands or shores. It is more desirable to investigate the mediation effect of multiple mediators simultaneously. We propose a Wald-type test for the overall natural indirect effect targeting the multiple mediator problem. Kernel machine based approach is employed to account for potential interactions and nonlinearity among mediators. Under certain regularity conditions, we develop a simple variance estimator when comparing coefficients from different models. The proposed test is evaluated on data simulated from various underlying mediation pathway structures and demonstrates substantial gain in power when the effects are nonlinear. We also apply the proposed test on the Normal Aging Study to investigate the effect of smoking on lung function measures that mediated by the DNA methylation level on different regions over the whole genome.

email: *jcshen@umich.edu*

ADAPTIVE CONTRAST WEIGHTED LEARNING FOR MULTI-STAGE MULTI-TREATMENT DECISION-MAKING

Yebin Tao*, University of Michigan

Lu Wang, University of Michigan

Dynamic treatment regimes (DTRs) are sequential decision rules that focus simultaneously on treatment individualization and adaptation over time. To directly identify the optimal DTR, we propose a

dynamic statistical learning method, adaptive contrast weighted learning, which can handle two or more treatment options at each stage. We develop semiparametric regression-based contrasts with the adaptation of treatment effect ordering for each patient at each stage, and the adaptive contrasts simplify the multi-treatment comparison problem to a weighted classification problem that can be solved by existing machine learning techniques. The algorithm is implemented recursively from the last stage, and we incorporate Q-learning for the backward induction. By combining semiparametric regression methods with machine learning algorithms, the proposed method is robust and efficient for the identification of the optimal DTR, as shown in the simulation studies. We illustrate our method using observational data on esophageal cancer.

email: *yebintao@umich.edu*

65. NEXT GENERATION SEQUENCING

SHRINKAGE OF DISPERSION PARAMETERS IN THE BINOMIAL FAMILY, WITH APPLICATION TO DIFFERENTIAL EXON SKIPPING

Sean Ruddy*, University of California, Berkeley

Marla Johnson, University of California, Berkeley

Elizabeth Purdom, University of California, Berkeley

The prevalence of sequencing experiments in genomics has led to an increased use of methods for count data in analyzing high-throughput genomic data. The importance of shrinkage methods in improving the performance of statistical methods remains. An example is gene expression data, where the counts per-gene are often modeled as an over-dispersed Poisson. Shrinkage estimates of the per-gene dispersion parameter have led to improved estimation of dispersion, particularly in the case of low sample sizes. We address a different count setting: comparing differential proportional usage via an over-dispersed binomial model. We are motivated by our interest in testing for differential exon skipping in mRNA-Seq experiments. We introduce a novel shrinkage method that models the over-dispersion with the double binomial distribution proposed by Efron (1986). Our method (WEB-Seq) is an empirical Bayes strategy for producing a shrunken estimate of dispersion and effectively detects differential proportional usage, and has close ties to the weighted-likelihood strategy of edgeR developed for gene expression data (Robinson and Smyth, 2007). We analyze its behavior on simulated and real data sets and show our method is fast, powerful and gives accurate control of FDR comparedto alternative approaches. Our method is available as the R-package, DoubleExpSeq, on CRAN.

email: *sruddy17@gmail.com*

ACCOUNTING FOR STOCHASTIC DROPOUT EVENTS IN DETECTING DIFFERENTIAL GENE EXPRESSION USING SINGLE-CELL RNA-Seq DATA

Cheng Jia*, University of Pennsylvania

Mingyao Li, University of Pennsylvania

Nancy Zhang, University of Pennsylvania

Recent advances in RNA-Seq technology has enabled the profiling of the whole transcriptome of individual cells. However, due to the unique statistical characteristics presented by single-cell RNA-Seq experiments, direct application of models built from bulk RNA-Seq data can cause biased inferences or reduced power. One of the major drawbacks of current methods is the failure to properly address the sample-specific stochastic dropout events dependent on the underlying gene expression. Here we present a flexible hierarchical mixture model framework that explicitly accounts for the sample-specific and gene-specific dropout probabilities, which are estimated with an empirical Bayes approach from spike-in data. Existing methods are also troubled by the inability to incorporate covariates when testing for differential gene expression. Our framework allows flexible modeling of covariates through a linear model. EM algorithms are implemented to estimate the parameters of biological interest, and likelihood ratio tests are designed to test for their significance.

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

NEXT-PEAK: A PER-BASE REGRESSION MODEL FOR ChIP-Seq PEAK CALLING

Nak-Kyeong Kim*, Virginia Commonwealth University

We propose a per-base regression model with a kernel of the normal-exponential two-peak (NEXT-peak) density for calling peaks in ChIP-seq data. The proposed NEXT-peak kernel parallels the physical processes generating the empirical data. The strand-specific, per-base tag count is assumed to be sum of the tag counts from the protein binding and the tag counts from the noise process. Unlike the existing models, the NEXT-peak model estimates strength of binding by computing the mixing probabilities between signal and noise; it also assigns a standard error to an estimated binding location. The comparison study with existing programs on real ChIP-seq datasets (STAT1 and NRSF) demonstrates that the NEXT-peak model performs well both in calling peaks and locating them.

email: *nak-kyeong.kim@vcuhealth.org*

GENE-SET ANALYSIS VIA COMBINING P-VALUES IN RNA-Seq DATA

Yu-Chung Wei*, U.S. Food and Drug Administration

Ching-Wei Chang, U.S. Food and Drug Administration

Nysia I. George, U.S. Food and Drug Administration

RNA (transcriptome) sequencing (RNA-seq) has become an important technology in studies of gene expression analysis. Since the information generated from a set of genes provides more biological or functional interpretation than a single gene, statistical methods for gene-set analysis of RNA-seq studies are needed. In this work, a testing statistic that utilizes a meta-analysis approach to pool individual p-values of genes in a gene set by accounting for the correlation structure among genes will be evaluated/modified for RNA-Seq data. The effects of different correlation estimators on testing results will be also evaluated. Simulation studies will be used to assess the performance.

email: *weiyuchung@gmail.com*

A MODEL FOR PAIRED-MULTINOMIAL DATA AND ITS APPLICATION TO ANALYSIS OF DATA ON A TAXONOMIC TREE

Pixu Shi*, University of Pennsylvania

Hongzhe Li, University of Pennsylvania

In human microbiome studies, the sequencing reads data are often summarized as counts of bacterial taxa at various taxonomic levels represented as a taxonomic tree. In addition, repeated measurements of mcirobiome are often obtained to assess change of microbial composition over time or after certain treatment. Existing models for such count data are often restricted in modeling the covariance structure of the counts and cannot handle paired multinomial data. We propose a new probability distribution for paired multinomial count data, which allows flexible covariance structure of the count data and can be used to model repeated measured multivariate counts. Based on this new distribution, we develop a statistic to test the difference in compositions based on paired multivariate count data. We demonstrate the application of the test for analysis of count data observed on a taxonomic tree in order to test change of mcirobiome composition over time and to identify the subtrees with different subcompositions. Our simulation results indicate that proposed test has correct type 1 errors and increased power compared to some commonly used methods.

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

A NOVEL NORMALIZATION METHOD FOR TIME SERIES METAGENOMIC COUNT DATA

Lingling An*, University of Arizona

Zhenqiang Lu, University of Arizona

Meng Lu, University of Arizona

Dan Luo, University of Arizona

Recent and rapid advent of high-throughput sequencing technologies has greatly promoted the field of metagenomics, which studies the genetic materials of entire microbial communities. Metagenomics has wide applications in Human Health and Medical Sciences, Environmental Biology, and Biodefense. A main question in metagenomic studies is "whether and how the microbial communities differ". Normalizing metagenomic count data plays a critical role for comparative analysis as microbial samples are rarely collected with the same amount and contain various types of noises. Several normalization methods have been developed for static metagenomic data; however, no approach has been proposed for normalizing time-course metagenomic data that contain more information than the static does. Taking advantage of time dependence property a novel approach is proposed for normalizing temporal count data. Compared with other existing methods the new approach shows the best performance through comprehensive simulation studies and real data analysis.

email: *anling@email.arizona.edu*

HOMOLOGY CLUSTER DIFFERENTIAL EXPRESSION ANALYSIS FOR INTERSPECIES mRNA-Seq EXPERIMENTS

Jonathan A. Gelfond*, University of Texas Health Science Center, San Antonio

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

Ming-Hiu Chen, University of Connecticut

Sun Wei, Fred Hutchinson Cancer Center

Kaitlyn Lewis, University of Texas Health Science Center, San Antonio

Sean Kinahan, Trinity University

Matthew Hibbs, Trinity University

Rochelle Buffenstein, Calico Labs

There is an increasing demand for exploration of the transcriptomes of multiple species with extraordinary traits such as the naked-mole rat (NMR). The NMR is remarkable because of its longevity and resistance to developing cancer. It is of scientific interest to understand the molecular mechanisms that impart these traits, and RNA-sequencing experiments with comparator species can correlate transcriptome dynamics with these phenotypes. Comparing transcriptome differences requires a homology mapping of each

transcript in one species to transcript(s) within the other. Such mappings are necessary, especially if one species does not have well-annotated genome available. Current approaches for this type of analysis typically identify the best match for each transcript, but the best match analysis ignores the inherent risks of mismatch when there are multiple candidate transcripts with similar homology scores. We present a method that treats the set of homologs from a novel species as a cluster corresponding to a single gene in the reference species, and we compare the cluster-based approach to a conventional best-match analysis in both simulated data and a case study with NMR and mouse tissues. We demonstrate that the cluster-based approach as superior power to detect differential expression.

email: *gelfondjal@uthscsa.edu*

66. ORAL POSTERS: GENOMICS

66a. INVITED ORAL POSTER: STATISTICAL METHODS FOR SINGLE-CELL RNA-Seq

Rhonda Bacher, University of Wisconsin

Jeea Choi, University of Wisconsin

Keegan Korthauer, Dana-Farber Cancer Institute

Ning Leng, Morgridge Institute for Research

Li-Fang Chu, Morgridge Institute for Research

James A. Thomson, Morgridge Institute for Research

Ron Stewart, Morgridge Institute for Research

Christina Kendziorski*, University of Wisconsin

Single-cell RNA-sequencing (scRNA-seq) has emerged as a revolutionary tool that allows us to address scientific questions that were elusive just a few years ago. The scRNA-seq technology has already enabled critical insights into novel sub-populations, differentiation progression, embryonic development, cancer, and neural diversity. Unfortunately, in contrast to the great biological and technological progress that has been made, statistical methods are lacking. Specifically, most scRNA-seq experiments use data analysis methods developed for bulk RNA-seq. On the surface, doing so seems reasonable since the basic data structure (a matrix of expression levels for m transcripts in n samples) is the same in bulk and single-cell experiments and, indeed, for some types of analyses, methods from bulk readily apply. However, the single-cell technology introduces technical artifacts not observed in bulk and for many types of analyses if these artifacts are not accommodated, biological signals are obscured, and may be distorted. Our poster highlights a number of challenges we are addressing in scRNA-seq studies including normalization, adjusting for nuisance variation, and the identification of sub-populations and genes showing differential dynamics across conditions.

email: *kendzior@biostat.wisc.edu*

66b. INVITED ORAL POSTER: THE WIDESPREAD AND CRITICAL IMPACT OF SYSTEMATIC BIAS AND BATCH EFFECTS IN SINGLE-CELL RNA-Seq DATA

Stephanie C. Hicks, Harvard University

Mingxiang Teng, Harvard University

Rafael A. Irizarry*, Dana-Farber Cancer Institute and Harvard University

Single-cell RNA-Sequencing (scRNA-Seq) has become the most widely used high-throughput method for transcription profiling of individual cells. Systematic errors, including batch effects, have been widely reported as a major challenge in high-throughput technologies. Surprisingly, these issues have received minimal attention in published studies based on scRNA-Seq technology. We examined data from five published studies and found that systematic errors can explain a substantial percentage of observed cell-to-cell expression variability. Specifically, we found that the proportion of genes reported as expressed explains a substantial part of observed variability and that this quantity varies systematically across experimental batches. Furthermore, we found that the implemented experimental designs confounded outcomes of interest with batch effects, a design that can bring into question some of the conclusions of these studies. Finally, we propose a simple experimental design that can ameliorate the effect of these systematic errors have on downstream results.

email: *rafa@jimmy.harvard.edu*

66c. CHANGE IN VARIANCE OF IGF2 GENE METHYLATION IS ASSOCIATED WITH THREE METABOLITES

Emily C. Hector*, University of Michigan

Jaclyn M. Goodrich, University of Michigan

Lu Tang, University of Michigan

Wei Perng, University of Michigan

Dana C. Dolinoy, University of Michigan

Adriana Mercado-Garcia, National Institute of Public Health, Mexico

Howard Hu, University of Toronto

Martha Maria Tellez-Rojo, National Institute of Public Health, Mexico

Karen E. Peterson, University of Michigan and Harvard School of Public Health

Peter X.K. Song, University of Michigan

Epigenetic mechanisms may mediate the association between early-life exposure to lead and metabolic derangements later in life. New access to high-dimensional longitudinal data has the potential to improve our understanding of complex relationships between early-life exposures, epigenetics, and adolescent metabolic outcomes. Analyzing epigenetic change over time as a predictor of phenotype is a relatively new area of study, and analysis typically focuses on change over time in mean percent methylation. Introducing a variance term as a predictor enables us to quantify how the variability in methylation over time, which may represent differences in regulatory control, predicts an outcome. Using data from the Early Life Exposures in Mexico to Environmental Toxicants cohort, we examined associations between change in methylation (Dmeth) and squared change in methylation (Dmeth2) between birth (n=70) and adolescence (n=250) at three genes and LINE-1 repetitive elements with three metabolites, xylose, X5.oxoproline, and C9.H2.N.P.S2, selected for their association with prenatal lead exposure. Random intercept models predicting concentrations of the three metabolites using Dmeth and Dmeth2, adjusting for age and sex, yielded significant associations for theIGF2gene methylation at several sites, suggesting that change in variance of methylation inIGF2may impact metabolism of these three metabolites in adolescence.

email: *ehector@umich.edu*

66d. PRIORITIZING GENES BASED ON BAYES FACTOR

Hongyan Xu*, Georgia Regents University

Fengjiao Hu, Georgia Regents University

Duchwan Ryu, Northern Illinois University

Varghese George, Georgia Regents University

In genomic analysis, such as differential gene expression analysis with microarray or RNA-seq, a list of genes will become significant from the initial statistical tests. In subsequent analysis, we are often faced with the problem of gene prioritization because only of a few genes can be followed up. Most of the gene prioritization methods are based on test p-values. However, this may not be appropriate because p-values does not reflect the probability of the alternative hypothesis. In this study, we propose a Bayesian approach for gene prioritization. Given different prior probabilities for the null and alternative hypothesis, Bayes factors can be calculated, which reflect the relative probability of the null and alternative hypothesis. We then rank genes according to the Bayes factors. Simulation results show that this approach can prioritizing genes with less false positives with varying cut-off values for test significance. This method can also incorporate prior information on the different hypotheses, which may be obtained from previous studies.

email: *hxu@gru.edu*

66e. RefCNV: IDENTIFICATION OF GENE-BASED COPY NUMBER VARIANTS USING WHOLE EXOME SEQUENCING

Lun-Ching Chang*, National Cancer Institute, National Institutes of Health

Biswajit Das, Leidos Biomedical Research Inc.

Chih-Jian Lih, Leidos Biomedical Research Inc.

Corrine Camalier, Leidos Biomedical Research Inc.

Paul McGregor, Leidos Biomedical Research Inc.

Eric Polley, National Cancer Institute, National Institutes of Health

With rapid advances in DNA sequencing technologies, whole exome sequencing (WES) has become a popular approach for detecting somatic mutations in oncology studies. The initial intent of the WES was to characterize single nucleotide variants and short insertions and deletions, but it was observed that the number of sequencing reads that mapped to a genomic region correlated with the DNA copy number. Numerous methods for the normalization and copy number variant identification have been proposed that make use of the observation that read coverage correlates with copy number, but most current methods depend on a matched germline sample. We propose a method RefCNV that does not require a matched normal sample, but instead we built a reference set that is used to estimate the distribution of the coverage for each exon. The construction of the reference set includes an evaluation of the sources of variability in the coverage distribution. We observed that sample processing steps had an impact on the distribution of the coverage, and therefore recommended using the same sample processing, sequencing, and bioinformatics steps in the reference set as those used on prospective samples. For each exon in the tumor sample, we compared the observed coverage with the expected normal coverage from the reference set. Thresholds for determining copy number variants were selected to control the false positive error rate. We presented examples of 13 cancer cell lines with known gene copy number variants on genesMET(7q31),EGFR(7p12) orERBB2(17q12). Three genes CNV results called by this algorithm correlated significantly with copy number detection using digital droplet PCR in 13 well-characterized cell lines.

email: *lun-ching.chang@nih.gov*

66f. A GENERALIZED FUNCTIONAL MODEL FOR ASSOCIATION ANALYSIS OF FAMILY-BASED SEQUENCING DATA

Sneha Jadhav*, Michigan State University

Qing Lu, Michigan State University

Family-based sequencing studies have been increasingly used in genetic research of complex diseases. While these studies hold

great promise for identifying new sequencing variants predisposing to complex diseases, the complex genetic structure of the sequencing data and various pedigree structures presented in the data pose statistical challenges on the association analysis. Functional linear models have been used with success to test for association between genetic variants and disease phenotypes. However, with the assumptions of independent samples and continuous phenotypes, it cannot directly apply to family-based association data, especially those with binary disease phenotypes. We therefore extend this framework and propose a generalized functional model (GFM) for association analysis of family data. We conducted simulation studies to investigate the type I error and the power of GFM. In the real data application, we applied GFM to a large-scale family-based sequencing data, evaluating the association of several candidate genes with nicotine dependence.

email: *jadhavsn@stt.msu.edu*

66g. BAYESIAN HIERARCHICAL MODELING AND SHRINKAGE PRIORS FOR GWAS

LiJin Joo*, New York University

Cheongeun Oh, New York University

We propose a Bayesian hierarchical model using shrinkage priors for fine mapping of Genome-wide Association Study (GWAS). A goal of GWAS is identifying genetic loci that are associated with a disease and fine mapping is a procedure that determines exact causal variants after candidate loci are proposed. A Bayesian approach provides flexibility to handle strong correlations among variables, called linkage disequilibrium and to estimate variants with rare frequencies. In order to model correlated variables, we introduce groups of variables and assign priors on groups pursuing group-wise shrinkages. In addition, the use of shrinkage priors, in a form of global-local mixtures, makes computation more efficient. In this study, we will investigate how the estimated scales influence on the accuracy of variable selection at gene-level and at variant-level. We analyzed the simulated data based on coalescent theory and the real data of hypertension in Hispanic population.

email: *lijin.joo@nyu.edu*

66h. STATISTICAL METHODS FOR COMPOSITIONAL DATA ANALYSIS WITH APPLICATION IN METAGENOMICS

Hongmei Jiang*, Northwestern University

Metagenomics is a powerful tool to study the microbial organisms living in a natural (such as soil, water) or host-associated (such as human gut) environment. Characterization of the relative abundance of the microbial organisms and their interactions is essential for understanding the structure of a microbial community and how the organisms live and work together as a community. It is well known that applying conventional statistical methods on compositional data may lead to biased results. In this talk we will discuss some novel methods for compositional data analysis including interaction patterns with application in metagenomics.

email: *hongmei@northwestern.edu*

66i. PATHWAY-BASED INTEGRATIVE BAYESIAN MODELING OF MULTI-PLATFORM GENOMICS DATA

Elizabeth J. McGuffey*, United States Naval Academy

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

Ganiraju C. Manyam, University of Texas MD Anderson Cancer Center

Raymond J. Carroll, Texas A&M University

Veerabhadran Baladandayuthapani, University of Texas MD Anderson Cancer Center

The identification of gene pathways involved in cancer development and progression and characterization of their activity in terms of multi-platform genomics can provide information leading to discovery of new targeted medications. Such drugs have the potential to be used for precision therapy strategies that personalize treatment based on the specific biology underlying an individual patient's cancer. We propose a two-step model that integrates multiple genomic platforms, as well as gene pathway membership information, to efficiently andsimultaneously (1) identify the genes significantly related to a clinical outcome, (2) identify the genomic platform(s) regulating each important gene, and (3) rank the pathways by importance to clinical outcome. We utilize a hierarchical Bayesian model with multiple levels of shrinkage priors to achieve efficient estimation, and our integrative framework allows us not only to identify the important pathways and the important genes within pathways, but also to gain insight as to the platform(s) driving the effects mechanistically. We apply our method to a subset of The Cancer Genome Atlas' publicly available glioblastoma multiforme data set and identify potential targets for future cancer therapies.

email: *emcguffe@usna.edu*

66j. INCORPORATING FUNCTIONAL 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

The bottlenecks in predictive capability of genome-wide association studies (GWAS) findings include unclear biological functions,

small to moderate effect sizes and small sample sizes. A recent study shows that regulatory variants are enriched with trait-associated single nucleotide polymorphisms (SNPs). In addition to collecting more samples, which is an expensive and time-consuming process, an alternative strategy that integrates biological knowledge with current GWAS data to improve risk prediction at individual level is appealing. In this study, we perform functional annotation in GWAS data and construct a weighted genetic prediction framework, which takes account of external biological information, for complex traits. Using simulations and real data analysis, we compare the weighted procedures to the standard procedures and to the best linear unbiased prediction based methods. We also investigate the impact of linkage disequilibrium on the prediction performance of predictive models.

email: *yue-ming.chen@uth.tmc.edu*

66k. DETECTION OF GENETIC INTERACTIONS THROUGH META-ANALYSIS AND EFFECT SIZE HETEROGENEITY

Yulun Liu*, University of Texas MD Anderson Cancer Center

Yong Chen, University of Pennsylvania School of Medicine

Paul Scheet, University of Texas MD Anderson Cancer Center

With varying, but substantial, proportions of heritability remaining unexplained by summaries of single-SNP genetic variation, there is a demand for methods that extract maximal information from genetic association studies. One source of variation that is difficult to assess is genetic interactions. A major challenge for naive detection methods is the large number of possible combinations, with a requisite need to correct for multiple testing. Assumptions of large marginal effects, to reduce the search space, may be restrictive and miss higher-order interactions with modest marginal effects. In this paper, we propose a new procedure for detecting gene-by-gene interactions through heterogeneity in estimated low-order (e.g. marginal) effect sizes by leveraging population structure, or ancestral differences, among studies in which the same phenotypes were measured. We implement this approach in a meta-analytic framework, which offers numerous advantages, such as robustness and computational efficiency, and is necessary when data-sharing limitations restrict joint analysis. We effectively apply a dimension reduction procedure that scales to allow searches for higher-order interactions. For comparison to our method, which we term phylogenY-aware Effect-size Tests for Interactions (YETI), we adapt an existing method that assumes interacting loci will exhibit strong marginal effects to our meta-analytic framework. As expected, YETI excels when multiple studies are from highly differentiated populations and maintains its superior- ity in these conditions even when marginal effects are small. When these conditions are less extreme, the advantage of our method wanes. We assess the Type-I error and power characteristics of complementary approaches to evaluate their strengths and limitations.

email: *Yulun.Liu@uth.tmc.edu*

66l. ASSESSING MITOCHONDRIAL DNA VARIATION AND COPY NUMBER USING TAILORED SEQUENCING ANALYSIS TOOLS

Jun Ding*, National Institute on Aging, National Institutes of Health

Carlo Sidore, Istituto di Ricerca Genetica e Biomedica, Consiglio Nazionale delle Ricerche, Italy

Francesco Cucca, Istituto di Ricerca Genetica e Biomedica, Consiglio Nazionale delle Ricerche, Italy

Goncalo R. Abecasis, University of Michigan

David Schlessinger, National Institute on Aging, National Institutes of Health

DNA sequencing identifies genetic variants for association studies, but studies typically focus on variants in nuclear DNA and ignore the mitochondrial genome. In fact, analyzing variants in mitochondrial DNA (mtDNA) presents special problems, which we resolve here with a general solution for the mtDNA analysis in sequencing studies. The new package comprises 1) an algorithm designed to identify mtDNA variants (i.e., homoplasmies and heteroplasmies), incorporating sequencing error rates at each base in a likelihood calculation and allowing allele fractions at a variant site to differ across individuals; and 2) an estimation of mtDNA copy number in a cell from whole-genome sequencing data. We also apply the methods to DNA sequence from lymphocytes of ~2,000 SardiNIA Project participants. Both homoplasmies and heteroplasmies show 5-fold higher transition/transversion ratios than variants in nuclear DNA. Also, heteroplasmy increases with age, though on average only ~1 heteroplasmy reaches the 4% level between ages 20 and 90. We find that mtDNA copy number averages ~110 copies/lymphocyte and is ~54% heritable, implying substantial genetic regulation of the level of mtDNA. To our knowledge, this is the largest population analysis to date of mtDNA dynamics, revealing the age-imposed increase in heteroplasmy and the high heritability of copy number.

email: *Jun.Ding@nih.gov*

66m. BAYESIAN LATENT HIERARCHICAL MODEL FOR TRANSCRIPTOMIC META-ANALYSIS TO DETECT BIOMARKERS WITH CLUSTERED META-PATTERNS OF DIFFERENTIAL EXPRESSION SIGNALS

Zhiguang Huo, University of Pittsburgh

Chi Song*, The Ohio State University

George Tseng, University of Pittsburgh

Due to rapid development of high-throughput experimental techniques and fast dropping prices, many transcriptomic datasets have been generated and accumulated in the public domain. Meta-analysis combining multiple transcriptomic studies can increase statistical power to detect disease related biomarkers. In this presentation, we introduce a Bayesian latent hierarchical model based

on one-sided p-values from differential/association analysis to perform transcriptomic meta-analysis. The p-value based method is capable of combining data from different microarray and RNA-seq platforms, and the latent variables help quantify homogeneous and heterogeneous differential expression signals across studies. A tight clustering algorithm is applied to detected biomarkers to capture differential meta-patterns that are informative to guide further biological investigation. Simulations and two examples using a microarray dataset from metabolism related knockout mice and an RNA-seq dataset from HIV transgenic rats are used to demonstrate performance of the proposed method.

email: *song.1188@osu.edu*

67. NEW STATISTICAL DEVELOPMENTS FOR EMERGING CHALLENGES WITH COMPLEX DATA STRUCTURES IN OBSERVATIONAL STUDIES

JOINT MODELING OF LONGITUDINAL HEALTH PREDICTORS AND CROSS-SECTIONAL HEALTH OUTCOMES VIA MEAN AND VARIANCE TRAJECTORIES

Michael R. Elliott*, University of Michigan

Bei Jiang, University of Alberta

Naisyn Wang, University of Michigan

Mary Sammel, University of Pennsylvania

Joint modeling of longitudinal and outcome data has been an active area in recent years: joint models of continuous longitudinal with time-to-event data (Proust-Lima et al. 2014), joint modeling of count and binary data via hidden Markov models (Jackson 2011) and connecting longitudinal biomarker to disease outcomes via a joint latent variable model (Jones et al. 2011). Typically these models treat within-subject variability as a nuisance parameter. Elliott (2007) flipped this paradigm by developing models that assumed underlying "clusters" of within-subject variability were related to the health outcome of interest, while the subject-specific trajectories were treated entirely as nuisance, showing evidence that such variability contained considerable information about the outcome. Others (Elliott et al. 2012, Jiang et al. 2015) have developed joint models to link information from mean trajectories and residual variance to health outcomes of interest, via models that use either shared random-effects or shared latent classes to jointly model longitudinal data and binary disease outcomes. We consider an application to predict severe hot flashes using the hormone levels collected over time for women in menopausal transition.

email: *mrelliot@umich.edu*

SPATIAL MEASUREMENT ERROR AND CORRECTION BY SPATIAL SIMEX IN LINEAR REGRESSION MODELS WHEN USING PREDICTED AIR POLLUTION EXPOSURES

Stacey E. Alexeeff, Kaiser Permanente

Raymond J. Carroll, Texas A&M University

Brent A. Coull*, Harvard University

Spatial modeling of air-pollution exposures has become widespread in air pollution epidemiology research as a way to improve exposure assessment. However, there are key sources of exposure model uncertainty when air pollution is modeled, including estimation error and model misspecification. We examine the use of predicted air pollution levels in linear health effect models under a measurement error framework. For the prediction of air pollution exposures, we consider a universal kriging framework, which may include land use regression terms in the mean function and a spatial covariance structure for the residuals. We derive the bias induced by estimation error and by model misspecification in the exposure model, and we find that a misspecified exposure model can induce asymptotic bias in the effect estimate of air pollution on health. We propose a new spatial SIMEX procedure, and we demonstrate that the procedure has good performance in correcting this asymptotic bias. We use a bootstrap procedure to estimate the standard errors in the spatial SIMEX method. We illustrate the spatial SIMEX approach in a study of air pollution and birthweight in Massachusetts.

email: *bcoull@hsph.harvard.edu*

INTRINSIC EFFICIENCY AND MULTIPLE ROBUSTNESS IN LONGITUDINAL DATA ANALYSIS WITH DROPOUT

Peisong Han*, University of Waterloo

Intrinsic efficiency and multiple robustness are two superior properties in missing data analysis. We study them in longitudinal data analysis with dropout. The idea is to calibrate the missingness probability at each visit using observed historical data. We consider one working model for the missingness probability and multiple working models for the data distribution. Intrinsic efficiency guarantees that, when the missingness probability is correctly modeled, the multiple data distribution models, combined with the observed data prior to the end of the study, are optimally accommodated to maximize efficiency. The efficiency generally increases as the number of data distribution models does, except for where one data distribution model is correctly specified as well, in which all the proposed estimators attain the semiparametric efficiency bound. Multiple robustness ensures estimation consistency if the missingness probability model is misspecified but one data distribution model is correct.

email: *peisonghan@uwaterloo.ca*

MULTIPLE ROBUST FITTING OF A LOG-LINEAR MODEL

Andrea Rotnitzky*, Di Tella University

Thomas Richardson, University of Washington, Seattle

We consider estimation of a log-linear model for k discrete variables adjusting for high-dimensional covariates. Due to the curse of dimensionality, it is practically impossible to estimate the parameters indexing the model for the dependence of a given interaction on covariates in a way that is robust to mis-specification of the models for other interactions. In this talk we show that it is possible to partially disentangle the estimation of different interactions. Specifically, for a given interaction Î»_{C} we provide an estimator that is consistent and asymptotically normal (CAN) for it under the disjunctive semiparametric model that assumes that for some iâ˜C, p(i?V?{i}) is correctly specified, or equivalently that for some iâ˜C , all the models for interactions Î»_{D} such that iâ˜D are correct. Furthermore, our estimator is locally semi-parametric efficient at the intersection of all of the models defining the disjunctive model. We provide a convex estimating function that is guaranteed to have at most one maximum, thus ensuring that fitting algorithms will produce at most one fit (regardless of starting values). Finally, by constructing a hierarchy of estimators we can estimate an entire log-linear model conditional on covariates in a robust manner.

email: *arotnitzky@utdt.edu*

68. STATISTICAL INNOVATIONS OF MASSIVE GENOMIC DATA ANALYSIS

THE GENERALIZED HIGHER CRITICISM FOR TESTING SNP-SET EFFECTS IN GENETIC ASSOCIATION STUDIES

Ian Barnett, Harvard School of Public Health

Rajarshi Mukherjee, Stanford University

Xihong Lin*, Harvard School of Public Health

It is of substantial interest to study the effects of genes, genetic pathways, and networks on the risk of complex diseases. These genetic constructs each contain multiple SNPs, which are often correlated and function jointly, and might be large in number. However, only a sparse subset of SNPs in a genetic construct are generally associated with the disease of interest. In this paper, we propose the Generalized Higher Criticism (GHC) to test for the association between a SNP set and a disease outcome. The higher criticism is a test traditionally used in high dimensional signal detection settings when marginal test statistics are independent and the number of parameters is very large. However these assumptions do not always hold in genetic association studies, due to linkage disequilibrium among SNPs and the finite number of SNPs in a SNP set in each genetic construct. The proposed GHC overcomes the limitations of the higher criticism by allowing for arbitrary correlation structures among the SNPs in a SNP-set, while performing accurate analytic p-value calculations for any finite number of SNPs in the SNP-set. We obtain the detection boundary of the GHC test. We compared empirically using simulations the power of the GHC method with existing SNP-set tests over a range of genetic regions with varied correlation structures and signal sparsity. We apply the proposed methods to analyze the CGEM breast cancer genome-wide association study.

email: *xlin@hsph.harvard.edu*

LEVERAGING ALGORITHMS FOR LOGISTIC REGRESSION WITH MASSIVE DATA

Ping Ma*, University of Georgia

For massive data with super-large sample size, it is computationally infeasible to obtain maximum likelihood estimates for unknown parameters, especially when the estimators do not have closed-form solutions. In this talk, I will present fast leveraging algorithms to efficiently approximate the maximum likelihood estimates in logistic regression models with binary responses, one of the most commonly used models in practice for classification. I will aslo present some theoretcial resulsts on consistency and asymptotic normality of the estimators. Synthetic and real data sets are used to evaluate the practical performance of the proposed methods.

email: *pingma@uga.edu*

STATISTICAL MODELING OF HIGH-THROUGHPUT RNA STRUCTURE PROBING DATA

Zhengqing Ouyang*, The Jackson Laboratory for Genomic Medicine

RNA structure has important roles in almost every step of RNA processing, including transcription, splicing, degradation, localization, and translation. Next generation sequencing technologies have emerged to dissect RNA structure at unprecedented levels. However, it remains challenging to analyze high-throughput RNA structure probing data because of various statistical issues. We present a novel statistical framework, named joint Poisson-gamma mixture (JPGM), for modeling high-throughput RNA structure probing data. Combining JPGM with hidden Markov models allows for the inference of RNA structure states at the genome scale and at single-nucleotide resolution. We apply it to various applications including both simulated and real datasets.

email: *zhengqing.ouyang@jax.org*

EXPANSION OF BIOLOGICAL PATHWAYS BY INTEGRATING ENORMOUS mRNA EXPRESSION DATASETS

Yang Li, Harvard University

Jun Liu*, Harvard University

Vamsi Mootha, Harvard Medical School

The number of publicly available gene expression datasets has been growing dramatically. Various methods had been proposed to predict gene co-expression by integrating the publicly available datasets. These methods assume that the genes in the query gene set are homogeneously correlated and consider no gene-specific correlation tendencies, no background intra-experimental correlations, and no quality variations of different experiments. We propose a two-step algorithm called CLIC (CLustering by Inferred Co-expression) based on a coherent Bayesian model to overcome these limitations. CLIC first employs a Bayesian partition model with feature selection to partition the gene set into disjoint co-expression modules (CEMs), simultaneously assigning posterior probability of selection to each dataset. In the second step, CLIC expands each CEM by scanning the whole reference genome for candidate genes that were not in the input gene set but co-expressed with the genes in this CEM. CLIC is capable of integrating over thousands of gene expression datasets to achieve much higher coexpression prediction accuracy compared to traditional co-expression methods. Application of CLIC to ~1000 annotated human pathways and ~6000 poorly characterized human genes reveals new components of some well-studied pathways and provides strong functional predictions for some poorly characterized genes. We validated the predicted association between protein C7orf55 and ATP synthase assembly using CRISPR knock-out assays. Based on the joint work with Yang Li and the Vamsi Mootha lab.

email: *jliu@stat.harvard.edu*

69. POLICY IMPLICATIONS OF SCIENTIFIC REPRODUCIBILITY - A PANEL DISCUSSION

DISCUSSANTS:

Constantine Gatsonis, Brown University

Marcia McNutt, Science (Editor-in-Chief)

Lawrence Tabak, National Institutes of Health (Principal Deputy Director)

Steven Goodman, Stanford University

70. MULTIVARIATE MODELS FOR SPATIALLY CORRELATED DATA

BAYESIAN MATRIX MODELS FOR MULTIVARIATE DISEASE MAPPING

Miguel A. Martinez-Beneito, Public Health Research Center of Valencia

Paloma Botella-Rocamora, CEU Cardinal Herrera University, Spain

Sudipto Banerjee*, University of California, Los Angeles

Multivariate disease mapping enriches traditional disease mapping studies by analyzing several diseases jointly. This enables one disease to borrow information from the others and produce improved estimates of the geographical distribution of their risks. Beyond multivariate smoothing for several diseases, several other factors, such as sex, age group, race, time period, and so on, could also be jointly considered to derive multivariate estimates. The resulting multivariate structures customarily induce an appropriate covariance model for the data. We introduce a formal framework for the analysis of multivariate data arising from the combination of more than two factors (geographical units and at least two more factors), what we have called Multidimensional Disease Mapping. We consider a rich and diverse class of models that subsume both separable and non-separable dependence structures and illustrate its performance on the study of real mortality data in Comunitat Valenciana (Spain).

email: *sudipto@ucla.edu*

MULTIVARIATE GENERALIZED LINEAR MODELS FOR SPACE-TIME DISEASE MAPPING

Marie Denis*, CIRAD, France

Sabastien Tisne, CIRAD, France

Indra Syahputra, PT Socfindo, Indonesia

Hubert de Franqueville, PalmElit SAS

Benoit Cochard, PalmElit SAS

In the field of epidemiology, a common objective is to study the mapping of diseases in relation to time and space. The conditional autoregressive (CAR) model is the most popular approach allowing flexible modeling of the spatial dependence structure. Hierarchical modeling is another commonly used approach that provides a flexible and effective tool for disease mapping. Based on these two approaches, we propose generalized linear models that incorporate a latent process for the spread of the disease. To model the dynamics of the disease in space and time, information from past observations in the area of interest and/or other areas are directly integrated in the process. We investigate the use of different neigh-

borhood structures and compare different models implemented with the integrated nested Laplace approximation. We illustrate the different models with an application to the spread of infection in oil palm trees.

email: *marie.denis@cirad.fr*

MULTIVARIATE LATENT STRUCTURE IN BAYESIAN SPATIO-TEMPORAL HEALTH MODELS

Andrew B. Lawson, Medical University of South Carolina

Mehreteab Aregay*, Medical University of South Carolina

Often geospatial disease outcomes can be profitably modelled together and their joint modeling leads to additional information concerning common etiology and shared confounding. With the addition of a temporal dimension this latent structural commonality can have time and space varying behavior. In addition there can be both joint and conditional dependence in the latent components. In this talk I focus on the possibility of the existence of temporal and spatial dependent scale ffects. An application to monitoring of a suite of weekly acute respiratory infection discharges within counties of South Carolina is considered and the latent switching of states is modelled.

email: *lawsonab@musc.edu*

A HIERARCHICAL BAYESIAN MODEL FOR PREDICTION OF MULTIVARIATE NON-GAUSSIAN RANDOM FIELDS

Frederic Mortier*, CIRAD, France

Pierrette Chagneau, INSA de Rennes, France

Nicolas Picard, Food and Agriculture Organization of the United Nations

Spatial mapping methods must be able to deal with multivariate and heterogeneous data, as most georeferenced datasets have these characteristics. In this talk, we present a hierarchical Bayesian approach based on spatial generalized linear mixed models, which allows the simultaneous modeling of dependent Gaussian, count and ordinal spatial fields. We use a moving average approach to model the spatial dependence between the processes. We show that this multivariate spatial hierarchical model has a superior predictive performance compared to univariate models. We illustrate its application on a real dataset collected in French Guiana to predict topsoil patterns.

email: *fmortier@cirad.fr*

71. METHODS FOR COMPARATIVE EFFECTIVENESS RESEARCH USING ELECTRONIC HEALTH RECORDS

COMPARING COMPARATIVE EFFECTIVENESS STUDIES USING ELECTRONIC HEALTH RECORD (EHR) DATA

Ruth Etzioni*, Fred Hutchinson Cancer Research Center

Lurdes Inoue, University of Washington

We consider comparing studies with a longitudinal biomarker and a failure time outcome in the setting of competing risks. The application is active surveillance (AS) in prostate cancer. AS is conservative management of low-risk prostate cancer. In AS, PSA is measured serially along with regular biopsies.The event of interest is the first positive biopsy, i.e. worse than the time of diagnosis.The competing risk (CR) consists of initiation of treatment without a positive biopsy. Different AS studies have different study protocols and CRs. We consider EHR data from two AS studies in which the cumulative incidence of first positive biopsy (in the presence of the CR) is different and compare the incidence of first positive biopsy in the absence of the CR to see if it is more comparable. Our methods use Bayesian joint longitudinal and competing risks failure time models and develop posterior estimates of the conditional probability of the event of interest in the absence of the CR by study. The methods account for differences in demographics, PSA distributions and AS protocols across studies.

email: *retzioni@fhcrc.org*

COMPARATIVE EFFECTIVENESS OF DYNAMIC TREATMENT STRATEGIES: THE RENAISSANCE OF THE PARAMETRIC G-FORMULA

Miguel Hernan*, Harvard University

Causal questions about the comparative effectiveness and safety of health-related interventions are becoming increasingly complex. Decision makers are now often interested in the comparison of interventions that are sustained over time and that may be personalized according to the individuals' time-evolving characteristics. These dynamic treatment strategies cannot be adequately studied by using conventional analytic methods that were designed to compare "treatment" vs. "no treatment". The parametric g-formula was developed by Robins in 1986 with the explicit goal of comparing generalized treatment strategies sustained over time. However, despite its theoretical superiority over conventional methods, the parametric g-formula was rarely used for the next 25 years. Rather, the development of causal inference methods for longitudinal data with time-varying treatments focused on semiparametric approaches. In recent years, interest in the parametric g-formula is

growing and the number of its applications increasing. This talk will review the parametric g-formula, the conditions for its applicability, its practical advantages and disadvantages compared with semiparametric methods, and several real world implementations for comparative effectiveness research.

email: *miguel_hernan@post.harvard.edu*

ELECTRONIC HEALTH RECORDS AS EVIDENCE GENERATION TOOLS FOR MEDICAL DECISION MAKING

Marianthi Markatou*, University at Buffalo

With the goal of transforming the US clinical research enterprise, the Institute of Medicine has called for a learning healthcare systemto accelerate the (cost-effective) generation of new evidence directly from, and applicable to, patient care processes. In this talk, we will first discuss a number of important challenges associated with the use of electronic health record (EHR) data for generation of new evidence. These challenges include issues of completeness, accuracy, complexity of data, biases, time parametrization and others. We will then discuss the use of electronic health record data for intelligent phenotyping that links health science and care. Finally, we will address the state of the art in the use of EHRs for evidence generation for medical decision making.

email: *markatou@buffalo.edu*

METHODS FOR MISCLASSIFIED TIME TO EVENT OUTCOMES IN STUDIES USING EHR-DERIVED ENDPOINTS

Rebecca A. Hubbard*, University of Pennsylvania

Weiwei Zhu, Group Health Research Institute

Jessica Chubak, Group Health Research Institute

Estimates of the relationship between an outcome and an exposure are biased by imperfect ascertainment of the outcome of interest. In studies using data derived from electronic health records (EHRs), misclassification of outcomes is common and is often related to patient characteristics. For instance, patients with greater co-morbid disease burden may use the healthcare system more frequently making it more likely that EHR will contain a record of their diagnosis, possibly resulting in poorer outcome classification for healthier patients. Misclassification-adjusted estimators in the context of time to event outcomes are complex and have primarily focused on discrete time proportional hazards models, which may not be appropriate for all study designs. Motivated by an algorithm for identifying breast cancer recurrence using EHR data, we investigated the implications of using an imperfectly assessed outcome in time-to-event analyses. We used simulation studies to demonstrate the magnitude of bias induced by failure to account for error in the status and timing of recurrence and compared alternative methods for correcting this bias. We conclude with general guidance on preferred methods to account for outcome misclassification in time to event studies using EHR data.

email: *rhubb@upenn.edu*

72. MISSING DATA ISSUES IN META-ANALYSIS WITH INDIVIDUAL PARTICIPANT DATA

MISSING CONFOUNDER DATA IN OBSERVATIONAL META-ANALYSIS WITH SYSTEMATICALLY MISSING DATA

Ian R. White*, MRC Biostatistics Unit, Cambridge, UK

Matthieu Resche-Rigon, Universite Paris Diderot

Individual participant data meta-analyses provide an excellent way to explore risk factor - disease associations and to construct prognostic models. Commonly, some covariates are completely unrecorded in some studies: we call this “systematically missing data”. This has been tackled in two ways. First, fully and partly adjusted associations can be combined in a multivariate meta-analysis. Secondly, systematically missing data can be multiply imputed. The latter requires suitable multilevel imputation models. The talk will briefly review the field, describe some new developments in multilevel imputation, and illustrate the methods using 12 cohort studies exploring risk factors for short-term mortality in acute heart failure.

email: *ian.white@mrc-bsu.cam.ac.uk*

ALLOWING FOR UNCERTAINTY DUE TO MISSING OUTCOME DATA IN PAIRWISE AND NETWORK META-ANALYSIS

Dimitris Mavridis*, University of Ioannina, Greece

Ian R. White, MRC Biostatistics Unit, Cambridge, UK

Julian PT. Higgins, University of Bristol, UK

Andrea Cipriani, University of Oxford, UK

Anna Chaimani, University of Ioannina, Greece

Georgia Salanti, University of Ioannina, Greece

Missing outcome data may compromise results of individual trials and their meta-analysis by potentially introducing bias in the estimated treatment effects. We propose a pattern-mixture model to estimate meta-analytic treatment effects for dichotomous and continuous outcomes when these are missing for some of the randomized individuals. The outcome in the missing participants is related to the outcome in the observed participants through an informative missingness parameter (IMP). In the absence of that parameter, data are missing at random (MAR). By informing the IMP (expert opinion or sensitivity analysis), we explore how robust

study and summary estimates are to departures from the MAR assumption. This is a two-stage method where in the first method we get an adjusted effect size and its standard error either using a Taylor series approximation or Monte Carlo while in the second stage the adjusted effect sizes are meta-analyzed using inverse variance. This method gives relatively less weight to studies with high missing rates to reflect the fact that there is uncertainty regarding the missing individuals. We developed Stata routines to employ the suggested methodology in pairwise and network meta-analyses. We illustrate the use of these routines using datasets from mental health.

email: *dimi.mavridis@gmail.com*

MULTIPLE IMPUTATION FOR HARMONIZING LONGITUDINAL NON-COMMENSURATE MEASURES IN INDIVIDUAL PARTICIPANT DATA META-ANALYSIS

Juned Siddique*, Northwestern University

Jerome P. Reiter, Duke University

Ahnalee Brincks, University of Miami

Robert D. Gibbons, University of Chicago

Catherine M. Crespi, University of California, Los Angeles

C. H. Brown, Northwestern University

There are many advantages to individual participant data meta-analysis for combining data from multiple studies. These advantages include greater power to detect effects, increased sample heterogeneity, and the ability to perform more sophisticated analyses than meta-analyses that rely on published results. However, a fundamental challenge is that it is unlikely that variables of interest are measured the same way in all of the studies to be combined. We propose that this situation can be viewed as a missing data problem in which some outcomes are entirely missing within some trials, and use multiple imputation to fill in missing measurements. We apply our method to 5 longitudinal adolescent depression trials where 4 studies used one depression measure and the fifth study used a different depression measure. None of the 5 studies contained both depression measures. We describe a multiple imputation approach for filling in missing depression measures that makes use of external calibration studies in which both depression measures were used. We discuss some practical issues in developing the imputation model including taking into account treatment group and study. We present diagnostics for checking the fit of the imputation model and investigating whether external information is appropriately incorporated into the imputed values.

email: *siddique@northwestern.edu*

BAYESIAN INFERENCE FOR MULTIVARIATE META-REGRESSION WITH A PARTIALLY OBSERVED WITHIN-STUDY SAMPLE COVARIANCE MATRIX

Hui Yao, Ernst & Young

Sungduk Kim*, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

Ming-Hui Chen, University of Connecticut

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

Arvind Shah, Merck

Jianxin Lin, Merck

Multivariate meta-regression models are commonly used in settings where the response variable is naturally multidimensional. Such settings are common in cardiovascular and diabetes studies where the goal is to study cholesterol levels once a certain medication is given. In this setting, the natural multivariate endpoint is low density lipoprotein cholesterol (LDL-C), high density lipoprotein cholesterol (HDL-C), and triglycerides (TG) (LDL-C, HDL-C, TG). In this article, we examine study level (aggregate) multivariate meta-data from 26 Merck sponsored double-blind, randomized, active, or placebo-controlled clinical trials on adult patients with primary hypercholesterolemia. Our goal is to develop a methodology for carrying out Bayesian inference for multivariate meta-regression models with study level data when the within-study sample covariance matrix S for the multivariate response data is partially observed. Specifically, the proposed methodology is based on postulating a multivariate random effects regression model with an unknown within-study covariance matrix âˆ‘ in which we treat the within-study sample correlations as missing data, the standard deviations of the within-study sample covariance matrix S are assumed observed, and given âˆ‘, S follows a Wishart distribution. Thus, we treat the off-diagonal elements of S as missing data, and these missing elements are sampled from the appropriate full conditional distribution in a Markov chain Monte Carlo (MCMC) sampling scheme via a novel transformation based on partial correlations. We further propose several structures (models) for âˆ‘, which allow for borrowing strength across different treatment arms and trials. The proposed methodology is assessed using simulated as well as real data, and the results are shown to be quite promising.

email: *kims2@mail.nih.gov*

73. MODELING HIGH DIMENSIONAL SPACE-TIME DATA WITH APPLICATIONS TO NEUROIMAGING

ESTIMATING INFORMATION FLOW IN LARGE BRAIN NETWORKS VIA CONVEX OPTIMIZATION

Xi Luo*, Brown University

Yi Zhao, Brown University

The brain can be conceptualized as a dynamic network of connected nodes, and information, such as external stimuli, is processed while passing through series of nodes that form pathways. This provides a foundational model for the spatial-temporal processes in the brain. This talk uses functional MRI data to study the problem of estimating information flow in large brain networks under event-related stimuli. We model such information flow as dynamic weights on each edge of the network. One challenge is that the number of pathways between the "source" and "target" nodes grows exponentially with the number of the nodes in the network. To this challenge, we develop a large-scale structural equation model, and we propose a constrained convex optimization approach to infer the model parameters. Our approach enjoys the following advantages. It relaxes the original non-convex "pathway-search" problem to a convex one with an innovation on a new penalty formulation that selects the major pathways. It incorporates various constraints that reflect the local conservation laws, inspired by the problem of studying fluid flows. The numerical merits are illustrated using simulated data and a real fMRI dataset.

email: *xi.rossi.luo@gmail.com*

A SCALABLE MULTI-RESOLUTION MODEL FOR ACTIVATION AND BRAIN CONNECTIVITY IN fMRI DATA

Stefano Castruccio*, Newcastle University

Hernando Ombao, University of California, Irvine

Marc Genton, King Abdullah University of Science and Technology, Saudi Arabia

Modeling the spatial dependence of fMRI data is an instrumental task to test the significance of local neurological activity and is one of the main challenges of contemporary neuroscience. For the sake of feasibility, standard models typically reduce dimensionality by modeling covariance between regions of interest - which are coarser or larger spatial units - rather than between voxels. However, ignoring this could drastically reduce our ability to detect activation patterns in the brain and hence produce misleading results. To overcome these problems we introduce a multi-resolution spatio-temporal model and a computationally efficient methodology to estimate cognitive control related activation and whole-brain connectivity. The proposed model allows to test for voxel-specific activation while accounting for non-stationary local spatial dependence within anatomically defined ROIs, as well as global (between-ROIs) dependence. The model is used in a motor-task fMRI study to investigate brain activation and connectivity patterns with the ultimate goal of finding associations between these patterns and regaining motor functionality following a stroke, using a single-subject fMRI data with 150,000 voxels per time frame, for a total of 22 million data points, using a high performance cluster to parallelize the inference.

email: *stefano.castruccio@ncl.ac.uk*

A NOVEL MULTISCALE METHODOLOGY FOR MULTIMODAL DATA INTEGRATION

John Aston, University of Cambridge

Jean-Marc Freyermuth*, University of Cambridge

Hernando Ombao, University of California, Irvine

In this talk we investigate the theoretical and practical properties of a general methodology for fusing complex spatio-temporal objects. These data typically exhibit strong anisotropic properties that our methodology can deal with particularly well. As a special case of application, we aim at combining EEG and fMRI data in a non-informed way so as to provide a basis for statistical inference benefiting from improved temporal and spatial resolutions.

email: *jmf84@cam.ac.uk*

ROBUST CLUSTERING METHODS FOR TIME-EVOLVING BRAIN SIGNALS

Tianbo Chen, King Abdullah University of Science and Technology, Saudi Arabia

Ying Sun*, King Abdullah University of Science and Technology, Saudi Arabia

Hernando Ombao, University of California, Irvine

Carolina Euan, Centro de Investigación en Matemáticas, Mexico

Brain activity following stimulus presentation and during resting state are often the result of highly coordinated responses of large numbers of neurons both locally (within each region) and globally (across different brain regions). Coordinated activity of neurons can give rise to oscillations which are captured by electroencephalograms (EEG). In this talk, new clustering methods will be presented for identifying synchronized brain regions where the EEGs show similar oscillations or waveforms, and the evolution of the identified clusters will be visualized dynamically. The method is developed for clustering EEG channels according to their spectral densities

estimated from multiple trials. The clustering algorithm is based on functional medians of the spectral density estimates, which are more robust compared to functional means when outliers are present. The medians are computed using notions of data depth designed for functional data, and different dissimilarity measures to compute the distance between spectral densities from any pairs of EEG channels are also explored and compared in the clustering algorithm. Our simulation studies suggest that the proposed method performs very well in producing the correct clusters even with introduced outlying signals. When applied to resting state EEG data, the method partly confirms the segmentation based on the anatomy of the cortical surface. In addition, we illustrate the dynamics of spectrally synchronized brain regions during resting state by visualizing the time-evolving clusters of the EEG channels in 3D environment. This talk is based on the joint work with Tianbo Chen, Hernando Ombao and Carolina Euan.

email: *ying.sun@kaust.edu.sa*

74. BAYESIAN HIERARCHICAL MODELING

BAYESIAN MIXED-EFFECTS VARYING-COEFFICIENT JOINT MODELS FOR SKEWED LONGITUDINAL DATA WITH APPLICATION TO AIDS CLINICAL STUDIES

Tao Lu*, State University of New York, Albany

In AIDS clinical study, two biomarkers, HIV viral load and CD4 cell counts, play important roles. It is well known that there is inverse relationship between the two. Nevertheless, the relationship is not constant but time varying. The mixed-effects varying-coefficient model is capable of capturing the time varying nature of such relationship from both population and individual perspective. In practice, the CD4 cell counts are usually measured with much noise and missing data often occur during the treatment. Furthermore, most of the statistical models assume symmetric distribution, such as normal, for the response variables. Often time, normality assumption does not hold in practice. Therefore, it is important to explore all these factors when modeling the real data. In this article, we establish a joint model that accounts for asymmetric distribution for the response variable, covariate measurement error and missingness simultaneously in the mixed-effects varying-coefficient modeling framework. A Bayesian inference procedure is developed to estimate the parameters in the joint model. The proposed model and method are applied to a real AIDS clinical study and various comparisons of a few models are performed.

email: *stat.lu11@gmail.com*

MODELLING PULSATILE HORMONE ASSOCIATIONS WITH COX CLUSTER MODELS

Huayu Liu*, University of Colorado, Anschutz Medical Campus

Nichole E. Carlson, University of Colorado, Anschutz Medical Campus

Alex J. Polotsky, University of Colorado, Anschutz Medical Campus

The negative effects of obesity on women's reproductive and offspring health are well-documented but the mechanisms remain poorly understood. Alterations in the secretion patterns of luteinizing hormone (LH) and follicle stimulating hormone (FSH) and their associations are hypothesized as potential mechanisms. These hormones are secreted intermittently in boluses, called pulses. The pulse release pattern and associations between the pulse releases of LH and FSH are the dominating regulatory mechanism for these hormones. Therefore, there is interest in modelling the pulse release process and associations of the processes between hormones. Here we show how a Cox cluster model of pulse locations can be used to quantify associations in the pulse secretion patterns. We embed the Cox cluster model into a joint model of hormone concentration profiles (the observed data). The Cox cluster model results in a more flexible association model for pulse location of two hormones compared to previous approaches that require co-occurrence of pulses. A spatial birth-death Markov chain Monte Carlo algorithm is used for estimation. Both simulation and experimental LH-FSH data are used to exhibit the performance of this model compared to existing methods.

email: *huayu.liu@ucdenver.edu*

A BAYESIAN FORMULATION FOR CAPTURING POPULATION HETEROGENEITY

Junxian Geng*, Florida State University

Elizabeth Slate, Florida State University

Population heterogeneity exists everywhere in real life. For instance, complexity of the underlying disease process may cause heterogeneity in the association between disease and biomarkers. In the context of binary markers such as single nucleotide polymorphisms (SNPs), we use ideas from logic regression and seek Boolean combinations that can explain association with a binary disease response. While we typically deal with binary data, our methods may also be used for continuous variables via appropriate discretization. We cast heterogeneity as unknown subgroups in the population; hence it is natural to adopt the Dirichlet process mixture model (DPMM) and mixture of finite mixture model (MFM) for our Bayesian formulation because of their clustering effect. We describe our model that incorporates the Boolean relations as parameters arising from a DPMM or MFM, and our way of addressing the associated challenges both in terms of specification of the base distribution and

estimation using a MCMC approach. For MCMC, we implement both an incremental algorithm (Gibbs sampler) and nonincremental algorithm (split and merge). We illustrate the performance of these methods with simulation and discuss applications.

email: *gengjunxianjohn@gmail.com*

BAYESIAN HIERARCHICAL MODELING TO DETERMINE SUBSTATE REPORTING AREAS

Tianyi Cai*, Harvard School of Public Health

Francesca Dominici, Harvard School of Public Health

Alan Zaslavsky, Harvard Medical School

Each year surveys are conducted to assess the quality of care for Medicare beneficiaries, using instruments from the Consumer Assessment of Healthcare Providers and Systems program. It is of interest to formulate a decision rule regarding the granularity of the reported survey results for Fee-for-Service beneficiaries. Depending on the heterogeneity of these results in each state, they can be presented pooled at the state level or unpooled at the substate level. In lieu of the current naïve hypothesis testing analysis, we propose and compare several Bayesian hierarchical models that combine information on small area means and variance components over states and over time, using data from 94 substate areas in 32 states from 2010 to 2014. We use estimates from our best-fitting models to identify the proper amount of pooling for presentation of direct estimates (state or substate level in each state) as well as to propose alternative small area estimates superior to either direct estimate. We further extend our model to include multiple outcomes, incorporating information contained in associations among the outcomes.

email: *cai01@fas.harvard.edu*

SPATIAL-TEMPORAL SURVIVAL ANALYSIS ON PROSTATE CANCER IN PENNSYLVANIA USING BAYESIAN ACCELERATED FAILURE TIME MODELS

Zheng Li*, Penn State College of Medicine

Ming Wang, Penn State College of Medicine

Stephen A. Matthews, Penn State Hershey Cancer Institute

Khaled Iskandarani, Penn State College of Medicine

Yimei Li, University of Pennsylvania

Vernon M. Chinchilli, Penn State College of Medicine

Prostate cancer is one of the most common cancers diagnosed among males, and is an important public health issue in Pennsylvania. The incidence rate and mortality vary substantially across geographical regions (counties) and over time (years). The widely-used Cox Proportional Hazards (PH) model does not apply due to the violation of proportional hazards assumption. In this work, we propose to use Bayesian accelerated failure time (AFT) models to analyze prostate cancer survivorship by incorporating random effects with multivariate conditional autoregressive (MCAR) priors for taking spatial temporal variation into account. The models are fitted based on Monte Carlo Markov Chain (MCMC) technique under the Bayesian framework. Extensive simulations are performed to examine and compare the performances of various Bayesian AFT models with MCAR priors. The criterion for model selection via the deviance information criterion (DIC) is also evaluated in the simulation study. Finally, we implement our method to the prostate cancer data obtained from the Pennsylvania Cancer Registry (PCR) which includes all reported prostate cancer diagnosed and death cases by county from years 2000-2011.

email: *zxl141@psu.edu*

HIERARCHICAL MULTIVARIATE SPACE-TIME METHODS FOR MODELING COUNTS WITH AN APPLICATION TO STROKE MORTALITY DATA

Harrison Quick*, Centers for Disease Control and Prevention

Lance A. Waller, Emory University

Michele Casper, Centers for Disease Control and Prevention

Geographic patterns in stroke mortality have been studied as far back as the 1960s, when a region of the southeastern United States became known as the “stroke belt” due to its unusually high rates of stroke mortality. While stroke mortality rates are known to increase exponentially with age, an investigation of spatiotemporal trends by age group at the county-level is daunting due to the preponderance of small population sizes and/or few stroke events by age group. Here, we harness the power of a complex, nonseparable multivariate space-time model which borrows strength across space, time, and age group to obtain reliable estimates of yearly county-level mortality rates from US counties between 1973 and 2013. Furthermore, we propose an alternative metric for measuring changes in event rates over time which accounts for the full trajectory of a county’s event rates, as opposed to simply comparing the rates at the beginning and end of the study period. In our analysis of the stroke data, we identify differing spatiotemporal trends in mortality rates across age groups, shed light on the gains achieved in the Deep South, and provide evidence that a separable model would be inappropriate for these data.

email: *harryq@gmail.com*

75. EPIDEMIOLOGIC METHODS

ACCOUNTING FOR INFORMED PRESENCE IN THE ANALYSIS OF ELECTRONIC HEALTH RECORDS

Benjamin A. Goldstein*, Duke University

Nrupen Bhavsar, Duke University

Matthew Phelan, Duke Clinical Research Institute

Michael J. Pencina, Duke University

Most observational analyses suffer from the potential for bias. In studies of electronic health records one serious source of potential bias is the fact that one's mere presence in the health record is itself informative. This can make inference challenging. Through a combined analysis of causal diagrams, simulations and observed data, I illustrate the potential for informed presence and how one can correct for it. Specifically I show the conditions under which adjusting for the number of encounters can help alleviate this bias. In doing so, I also illustrate the conditions under which this can result in residual confounding through M-Bias: bias from conditioning on a collider. Causal theory and analytic results show that the sensitivity of the disease diagnosis algorithm is inversely related to confounding due to informed presence and directly related to the potential for M-Bias. The results have implications for analysis of EHRs and administrative data.

email: *ben.goldstein@duke.edu*

SPATIAL PATTERNING OF DIABETES IN DURHAM, NORTH CAROLINA: A BAYESIAN ANALYSIS OF ASSOCIATIONS WITH INDIVIDUAL AND NEIGHBORHOOD CHARACTERISTICS

Mercedes A. Bravo*, Children's Environmental Health Initiative, University of Michigan

Rebecca Anthopolos, Children's Environmental Health Initiative, University of Michigan

Marie Lynn Miranda, Rice University

Background:We conducted spatial modeling to assess whether features of the neighborhood environment are associated with increased risk of incident diabetes.Methods:Individual-level data for Durham, North Carolina were obtained from the Duke University Health System (2007-2011). Patient data were linked to block-level racial isolation (RI) and built environment (BE) indices. We assessed spatial variation in diabetes and its relationship with neighborhood variables (e.g., RI, BE) using Bayesian hierarchical models with spatially structured intrinsic conditional autoregressive (ICAR) random effects.Results:In the null model, the posterior spatial variance of diabetes was 0.184 (95% Credible Interval: 0.143, 0.222), and 6% (89/1,377) of neighborhoods (blocks) had elevated risk. However, RI of non-Hispanic blacks accounted for a 17% reduction in the spatial variation of diabetes, leaving less than 3% of neighborhoods (37/1,377) with residual elevated risk. In comparison, controlling for standard individual-level risk factors resulted in 7 neighborhoods with residual increased risk. Poor quality BE was not associated with diabetes.Conclusion:We quantify spatial risk of diabetes diagnosis in Durham, NC, identify blocks with elevated risk, and evaluate the role of neighborhood variables in diabetes risk. Racial isolation is an easily computed neighborhood-level index that may be useful to identify at-risk neighborhoods.

email: *mbravo@med.umich.edu*

THE ASSOCIATIONS OF DRUGS WITH ACUTE MYOCARDIAL INFARCTION: BIAS CORRECTION, GLOBAL PROFILING AND INFERENCE ON INDIVIDUAL DRUG

Changyu Shen, Indiana University School of Medicine and School of Public Health

Xiaochun Li, Indiana University School of Medicine and School of Public Health

Jia Zhan*, Indiana University School of Medicine and School of Public Health

We address two issues in drug-outcome association studies. First, it has been recognized that electronic health records (EHR) databases may have hidden biases, for example, failure or incomplete capture of exposure and covariates, such that confounding cannot be fully controlled. Consequently, risk estimates may be biased, resulting in the misguided assessment of the strength and direction of drug-outcome associations. Second, the distribution of the risk measures of a large number of drugs on market for a given outcome is unknown. Using acute myocardial infarction (AMI) as an example, we illustrate how the first issue can be addressed by calibrating the risk measures through drugs known to have no association with AMI in a population-level electronic medical records database (the Indiana Network for Patient Care). We then employ an empirical Bayes approach to address the second issue, which helps to improve the accuracy for the inference of the association of an individual drug with AMI. The study shows that without the hidden bias correction, 66.5%, 12.1% and 2.6% of the drugs included have a risk ratio for AMI greater than 1, 1.5 and 2, respectively. After the hidden bias correction, the proportions become 50.8%, 7.4% and 1.7%, respectively. Using the empirical Bayes method, we gain 47% (without bias correction) and 49% (with bias correction) precision for the estimation of the risk ratio of a hypothetical new drug. Our approach serves as a general strategy for pharmaco-epidemiology studies for either an individual drug-outcome pair or multiple drug-outcome pairs.

email: *jiazhan@umail.iu.edu*

APPROXIMATE BAYESIAN COMPUTATION FOR COMPARTMENTAL EPIDEMIC MODELS - METHODS AND SOFTWARE

Grant D. Brown*, University of Iowa

Aaron T. Porter, Colorado School of Mines

Jacob J. Oleson, University of Iowa

Epidemic modeling techniques allow investigators to better understand the spread of diseases by quantifying pathogen behaviors, and allowing users to weigh the evidence for particular modes of transmission. These models also provide the ability to forecast future spread, suggest new public health interventions, and evaluate existing ones. Nevertheless, implementation of epidemic models can be difficult due to their complex nature and the presence of poor or missing data. F221We propose a general class of spatial SEIRS compartmental models in a hierarchical Bayesian framework, along with software designed to perform such analyses efficiently using Approximate Bayesian Computation via Sequential Monte Carlo (ABC-SMC). We will begin by introducing ABC techniques, followed by a brief introduction to the ABSEIR R package. Particular attention will be paid to the evaluation of spatial and intervention related hypotheses, using the example of endemic cholera spread in Haiti and the Dominican Republic.

email: *grant-brown@uiowa.edu*

PROPORTIONAL HAZARDS REGRESSION FOR INTERVAL-CENSORED FAILURE TIME DATA IN CASE-COHORT STUDIES

Qingning Zhou*, University of North Carolina, Chapel Hill

Haibo Zhou, University of North Carolina, Chapel Hill

Jianwen Cai, University of North Carolina, Chapel Hill

The case-cohort design has been commonly used to reduce costs of assembling or measuring expensive covariates in large cohort studies. The existing work on this design mainly focuses on right-censored data. In practice, however, the failure time is often subject to interval-censoring, that is, the failure time is never exactly observed but known only to fall within some random time interval. In this talk, we consider the case-cohort study design with interval-censoring and fit the proportional hazards model to data arising from this design. We employ the inverse probability weighted likelihood function and propose a sieve estimation approach via Bernstein polynomials. The consistency and asymptotic normality of the resulting regression parameter estimator are established and the weighted bootstrap procedure is suggested for variance estimation. Simulation results show that the proposed method works well for practical situations, and an application is provided for illustration.

email: *qz4z3@mail.missouri.edu*

BIAS AND ARTIFACT TRADE-OFF IN MODELING TEMPORAL TREND OF ARCHIVED DATA WITH APPLICATIONS TO PUBLIC HEALTH STUDIES, DEMOGRAPHY, MARKETING RESEARCH AND SOCIOLOGY

Martina Fu, Stanford University

David Todem, Michigan State University

Wenjiang Fu*, University of Houston

Shuangge Ma, Yale University

In public health studies, it is important to estimate accurately the temporal trend of disease incidence and mortality. Often the disease mortality rate varies with the age of patients, a summary rate is often estimated based on a sequence of age-specific rates. The approach is known to be complex because of the Simpson's paradox and because the age structure varies with time due to aging. The crude rate is well known unfair for comparison, a direct age-standardization method has been employed to calculate age-adjusted rate using the age-structure of a standard population (often the US year 2000 population). Although the same method has been applied to demography, economics, marketing research and sociology as a standard procedure, it has been criticized for the lack of justification. In this work, we will study this procedure, point out that it inevitably introduces bias, overestimates cancer mortality rates, but underestimates cancer case fatality rates. We provide an upper bound of such bias, and further point out that the crude rate is incomparable because of the artifact due to varying age structure. We introduce a novel mean reference population method for bias-artifact trade-off. It removes the artifact, minimizes the bias, and improves the estimation accuracy.

email: *fuw@math.uh.edu*

76. GWAS: APPLICATIONS

SHRINKAGE-BASED GENOME WIDE ASSOCIATION ANALYSES BASED ON SPARSE VERSUS GAUSSIAN PRIORS

Chunyu Chen*, Michigan State University

Juan P. Steibel, Michigan State University

Robert J. Tempelman, Michigan State University

Genomic best linear unbiased prediction (GBLUP) has been increasingly adapted for genome-wide association (GWA) analyses. A currently popular modification of GBLUP for GWA, which we label as classical GBLUP (c-GBLUP), is to specify all genetic marker effects as being Gaussian, a priori, except for the marker being tested, whereas shrinkage GBLUP (s-GBLUP) treats all marker effects as Gaussian. The c-GBLUP procedure has previously been demonstrated to preserve Type I error rates whereas s-GBLUP

seemingly leads to overly conservative GWA tests. Heavy-tailed (BayesA) or variable selection (SSVS) priors are sparser specifications that may provide more robust GWA testing. Given that MCMC techniques can be computationally onerous, we propose inferences under these alternative priors based using the EM algorithm. Using both simulation study and data from a F2 cross between Duroc and Pietrain pigs, we discovered that BayesA and SSVS shrink the majority of the posteriorz-score basedP-values to be larger relative to c-GBLUP, whereas markers in putative QTL regions tend to have substantially smallerP-values in BayesA and SSVS compared to c-GBLUP. We suggest that BayesA and SSVS or other sparse specifications provide promising alternatives for GWA analyses of complex traits, particularly where a global null hypothesis of null effects may not be particularly appropriate.

email: *chench57@msu.edu*

AN EXPOSURE-WEIGHTED SCORE TEST FOR GENETIC ASSOCIATIONS INTEGRATING ENVIRONMENTAL RISK FACTORS

Summer S. Han*, Stanford University

Philip S. Rosenberg, National Cancer Institute, National Institutes of Health

Arpita Ghosh, Public Health Foundation of India

Maria Teresa Landi, National Cancer Institute, National Institutes of Health

Neil E. Caporaso, National Cancer Institute, National Institutes of Health

Nilanjan Chatterjee, Johns Hopkins University

Current methods for detecting genetic associations lack full consideration of the background effects of environ- mental exposures. Recently proposed methods to account for environmental exposures have focused on logistic regressions with gene–environment interactions. In this report, we developed a test for genetic association, encompassing a broad range of risk models, including linear, logistic and probit, for specifying joint effects of genetic and environmental exposures. We obtained the test statistics by maximizing over a class of score tests, each of which involves modified standard tests of genetic association through a weight function. This weight function reflects the potential heterogeneity of the genetic effects by levels of environmental exposures under a particular model. Simulation studies demonstrate the robust power of these methods for detecting genetic associations under a wide range of scenarios. Applications of these methods are further illustrated using data from genome-wide association studies of type 2 diabetes with body mass index and of lung cancer risk with smoking.

email: *summer.han@stanford.edu*

DETECTING SHARED GENETIC VARIANTS BETWEEN TWO DISEASES WITH DEPENDENT SNPs

Wanjie Wang*, University of Pennsylvania

Tony Cai, University of Pennsylvania

Hongzhe Li, University of Pennsylvania

It is of great interest to detect the genetic associations between two closely related diseases, such as Type 2 diabetes and hypertension. The detection and identification of shared genetic variants may help to understand the genetic architecture for these diseases. Procedures (Zhao et al) are established to solve the this problem, by testing whether there are SNPs simultaneously associated to both diseases, assuming the SNPs are independent. We propose a more realistic latent variable model, where the high correlations between SNPs in each linkage disequilibrium (LD) block are addressed. Using this model, we propose a new method that inherits the principal component score and global maximum statistic to test the shared genetic variants between two diseases. This proposed method is easy to implement and fast to compute. For the p-value calculation, we propose an analytical expression under certain conditions, and a permutation procedure in real data analysis. In simulation studies, the proposed method performs better than other standard methods. In real data analysis, we apply it to data from 10 diseases and identify genetic associations for three pairs of these diseases.

email: *wanjiew@wharton.upenn.edu*

DETECTION OF SIGNAL REGIONS IN WHOLE GENOME ASSOCIATION STUDIES

Zilin Li*, Tsinghua University and Harvard School of Public Health

Xihong Lin, Harvard School of Public Health

We consider in this paper detection of signal regions associated with disease phentoypes in whole genome array and sequencing association studies. The existing gene- or region-based methods test for the associations of the genetic variants in pre-specified regions with phenotypes. We propose a quadratic scan statistic based method to detect the existence and the locations of signal segments by scanning the genome. The proposed method accounts for the correlation (linkage disequilibrium) among genetic variants, allowing for signal genetic variants which have effects in different directions and are mixed with neutral variants in signal regions. We derived an asymptotic threshold to control the familywise error rate, and show that, under regularity conditions, the proposed method consistently selects the true signal regions. We performed simulation studies to evaluate the finite sample performance of the proposed method. Our simulation results showed that the proposed procedure has a better finite sample performance compared to the existing methods, especially in the presence of variant effects in different directions,

or neutral variants and the correlation among variants in signal regions. We applied the proposed procedure to analyze a lung cancer genome-wide association study to identify the regions of SNPs which are associated with lung cancer risk.

email: *li@hsph.harvard.edu*

A COMPARISON STUDY OF FIXED AND MIXED EFFECT MODELS FOR GENE LEVEL ASSOCIATION STUDIES OF COMPLEX TRAITS

Chi-Yang Chiu*, 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

Jeesun Jung, National Institute on Alcohol Abuse and Alcoholism, National Institutes of Health

Daniel E. Weeks, University of Pittsburgh

Alexander F. Wilson, National Human Genome Research Institute, National Institutes of Health

Christopher I. Amos, Dartmouth Medical School

Joan E. Bailey-Wilson, National Human Genome Research Institute, National Institutes of Health

James L. Mills, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

In association studies of complex traits, fixed effect models are usually used to test for association between phenotypic traits and major gene loci. In recent years, variance-component tests based on mixed models were developed for region-based genetic variant association tests. In the mixed models, the association is tested by a null hypothesis of zero variance via a sequence kernel association test (SKAT) and its optimal unified test (SKAT-O). Although there are some comparison studies to evaluate the performance of mixed and fixed models, there is no systematic analysis to determine when the mixed models perform better and when the fixed models perform better. Here we evaluated, based on extensive simulations, the performance of the fixed and mixed model statistics for quantitative traits, using genetic variants located in 6 and 12 kb simulated regions. We found that the fixed effect tests have accurately controlled false positive rates. In most cases, either one or both of the fixed effect tests are more powerful or similar to the mixed models except for the case of all causal variants are rare and region size is 12 kb. We argue that the fixed effect models are useful in most cases.

email: *chiuchiyang@gmail.com*

THE CORRECTION OF CELL-TYPE COMPOSITION IN EPIGENOME-WIDE ASSOCIATION STUDIES

Shaoyu Li*, University of North Carolina, Charlotte

Exploring how DNA methylation variants, which are believed to play important roles in many fundamental biological processes, influence the disease outcome is of increasing interest in recent years. Accelerated by the flying high-throughput biotechnologies, genome scale systematic epigenomics analogous to GWASs, epigenome-wide association studies (EWASs) have become affordable and practical. However, EWASs are facing a major complicating factor that is not usually considered in GWASs: DNA methylation patterns are specific to different cell types. This could lead to false associations because the cell type mixture proportions in peripheral blood and tissue samples are usually different. We propose a Bayesian mixed-model to adjust the cell-type heterogeneity effect in EWASs and evaluate the performance of the method in terms of statistical characteristics and computational cost by comparison with the state-of-art approach.

email: *sli23@uncc.edu*

77. MISSING DATA

ESTIMATING THE MARGINAL EFFECT OF INTERVENTIONS TO REDUCE SPREAD OF COMMUNICABLE DISEASES: WHAT CAN BE GAINED FROM CONTACT NETWORK INFORMATION?

Melanie Prague*, Harvard School of Public Health

Patrick Staples, Harvard School of Public Health

JP Onnela, Harvard School of Public Health

Eric Tchetgen Tchetgen, Harvard School of Public Health

Victor De Gruttola, Harvard School of Public Health

We develop methods to leverage information on contact networks to improve efficiency and validity of analysis of data from cluster-randomized trials (CRTs) designed to investigate the impact of prevention interventions. These methods make use of flexible approaches for network generation that are based on degree-corrected stochastic block models and that are potentially applicable for a wide range of transmissible diseases. A semi-parametric doubly robust estimator is useful for estimating the marginal effect of the intervention while adjusting for imbalance in covariates and missing information. Simulations of CRTs show that adjusting for network features increases efficiency, in some cases, by more than 30% and leads to considerable increase of the statistical power while retaining coverage near the nominal value of 95%. We identify the major network features driving these improvements, such as the number of first- and second-degree contacts, the number of closed subcommunities, and the shortest path to an infected individual. Because network information is difficult to collect and networks are likely to be modeled with considerable uncertainty, we perform a sensitivity analysis evaluating the degree to which it is profitable to use partial

or approximate information. Adjustments for covariates measured with error based on regression calibration are investigated.

email: *mprague@hsph.harvard.edu*

A DOUBLE ROBUST SEMIPARAMETRIC METHOD TO ACCOUNT FOR MISSING CONFOUNDER DATA

Katherine L. Evans*, Harvard University

Eric Tchetgen Tchetgen, Harvard School of Public Health

Missing data and confounding are two problems researchers face in observational studies for comparative effectiveness. Williamson et al (2012) proposed a unified approach to handle both issues concurrently using a multiply-robust (MR) methodology for missing confounder information. We show that while their approach is MR in theory, there are implicit assumptions regarding model congeniality that are unlikely to hold in practice, which implies their approach will in fact fail to be multiply robust under a standard parametrization. To address this, we propose an alternative transparent parameterizations of the likelihood function, which makes explicit model dependencies between various nuisance functions needed to evaluate the efficient score. The proposed method is genuinely doubly-robust (DR) in that it is consistent and asymptotic normal if one of two sets of modeling assumptions holds, and we establish that in a sense, this is the best one can do in this framework, and that while MR remains theoretically possible, in practice the property will not hold exactly. We evaluate the performance of the DR method via simulation study and apply the method to assess the effect of surgical resection on survival time among patients with glioblastoma multiforme.

email: *kevans@fas.harvard.edu*

ON INVERSE PROBABILITY WEIGHTING FOR NONMONOTONE MISSING AT RANDOM DATA

BaoLuo Sun*, Harvard School of Public Health

Eric Tchetgen Tchetgen, Harvard School of Public Health

The development of coherent missing data models to account for nonmonotone missing at random (MAR) data by inverse probability weighting (IPW) remains to date largely unresolved. As a consequence, IPW has essentially been restricted for use only in monotone missing data settings. We propose a class of models for nonmonotone missing data mechanisms that spans the MAR model, while allowing the underlying full data law to remain unrestricted. For parametric specifications within the proposed class, we introduce an unconstrained maximum likelihood estimator for estimating the missing data probabilities which can be easily implemented using existing software. To circumvent potential convergence issues with this procedure, we also introduce a Bayesian constrained approach to estimate the missing data process which is guaranteed to yield inferences that respect all model restrictions. The efficiency of the standard IPW estimator is improved by incorporating information from incomplete cases through an augmented estimating equation which is optimal within a large class of estimating equations. We investigate the finite-sample properties of the proposed estimators in a simulation study and illustrate the new methodology in an application evaluating key correlates of preterm delivery for infants born to HIV infected mothers in Botswana, Africa.

email: *bluosun@gmail.com*

MAXIMUM LIKELIHOOD ESTIMATION IN A SEMICONTINUOUS REGRESSION MODEL WITH A COVARIATE SUBJECT TO A DETECTION LIMIT

Paul W. Bernhardt*, Villanova University

A semicontinuous regression model has a response that is continuous over certain regions but also has one or more point masses. While semicontinuous models have been studied in a variety of contexts, they have not been considered in a regression scenario where a covariate is subject to a detection limit. In the motivating Genetic and Inflammatory Markers of Sepsis study, one goal was to determine how biomarkers subject to detection limits are related to survival time for patients entering the hospital with community acquired pneumonia. Since patients surviving at least 90 days were consider cured, and no patients were lost to follow-up prior to 90 days, a usual survival model does not apply. We propose a two-part regression model for this situation where the probability of being cured is modeled using logistic regression and the distribution of events that occur is modeled using a truncated accelerated failure time model. To estimate the parameters in this model, we propose a Monte Carlo EM algorithm where the conditional expectations in the E-step are approximated by averaging over multiply generated values for the covariate subject to detection limits. We suggest using a mixture of normals to flexibly model the censored covariate.

email: *paul.bernhardt@villanova.edu*

FEASIBILITY OF VARIABLE-BY-VARIABLE IMPUTATION IN CLUSTERED DATA WITH MULTIPLE MEMBERSHIP

Tugba Akkaya-Hocagil*, State University of New York, Albany

Recai M. Yucel, State University of New York, Albany

This work concerns with incomplete data in structures with correlated observational units due to their multilevel and/or cross-classified nature (e.g. students nested within classrooms within schools, or patients cross-classified by hospitals and patients' residential neighborhoods). We also consider lowest level observational units to belong to multiple higher level clustering factors such as students attending multiple schools. We devise inference by multiple imputa-

tion to explicitly incorporate uncertainty due to missing data into the underlying inferences. Our multiple imputation routines make use of flexible variable-by-variable methods to entertain the potentially diverse set of variables as well as common survey practices such as skip patterns. In this presentation we discuss the feasibility of this method. Particularly, through comprehensive simulation study, we compare and contrast these methods with their joint counterparts using only the Gaussian variables with missing values.

email: takkayahocagil@albany.edu

MIXED-EFFECTS MODELS FOR MULTIVARIATE CLUSTERED DATA WITH NONIGNORABLE MISSING OUTCOMES

Jiebiao Wang*, University of Chicago

Pei Wang, Icahn Medical School at Mount Sinai

Lin S. Chen, University of Chicago

Multivariate clustered data are commonly collected in various fields. Jointly analyzing multiple outcomes takes into account the correlations among the outcomes, and may improve the power to detect the effect of a covariate on the outcomes. When there is substantial amount of missingness in the data, if the missing data are not missing at random, ignoring them may introduce bias and reduce efficiency. In this work, based on a general framework for clustered data with non-ignorable missing outcomes, we propose multivariate mixed-effects models to explicitly model the correlations among the outcomes and the missing data mechanisms. The modeled missingness can depend on the missing outcomes or the random effects. Based on simulation studies, we illustrate the advantages of our proposed methods over other methods that either ignoring the missing data or the relationships among the outcomes. We apply the proposed methods to analyze multiple peptides from each protein in a breast cancer proteomics dataset.

email: jwang88@uchicago.edu

A SIMPLE METHOD OF ESTIMATING THE ODDS RATIO WITH INCOMPLETE DATA IN 1:N MATCHED CASE-CONTROL STUDIES

Chan Jin, Augusta University

Stephen W. Looney*, Augusta University

A 1:n matched case-control design, in which each case is matched to n controls, is commonly used to control for confounders when estimating the exposure-disease (E-D) association. The odds ratio (OR) is typically used to quantify such an association. However, when the exposure status is unknown for at least one individual in a matched case-control grouping, difficulties in estimating the true OR may arise. If the exposure status is known for all individuals in each case-control grouping, conditional logistic regression is an effective method for estimating the true OR. When cases and controls are not matched, and their exposures can be assumed to be independent, the cross-product ratio from a single exposure-by-disease contingency table is an effective way to estimate the OR. In this presentation, we propose a simple method for estimating the OR when the sample consists of a combination of matched and unmatched cases and controls, which can result when there is incomplete 1:n matching. This method is based on a weighted average of traditional methods for estimating the OR with matched and unmatched case-control data. We use simulation to compare our method to existing methods for analyzing incompletely matched 1:n data.

email: slooney@gru.edu

78. SEMI-PARAMETRIC AND NON-PARAMETRIC SURVIVAL ANALYSIS

REGRESSION ANALYSIS OF CURRENT STATUS DATA WITH GENERALIZED ODDS-RATE HAZARDS MODELS

Bin Yao*, University of South Carolina

Lianming Wang, University of South Carolina

Generalized odds-rate hazards (GORH) models, also referred to as G rho family, are a general class of semiparametric regression models containing several popular survival models, such as the proportional hazards model and the proportional odds model. Although many approaches have been proposed for analyzing right-censored data using the GORH models, little research has been reported for analyzing current status data. In addition, most of the existing approaches with the GORH models assume rho is known because estimating rho together with other parameters are problematic. This article investigates the nonidentifiablity issues of the GORH models when treating rho as an unknown parameter. A computationally efficient EM algorithm is proposed for analyzing current status data when rho is known. The proposed approach is robust to initial values, fast to converge, and provides variance estimates in closed form. When rho is unknown, a direct generalization of our method allows estimating rho is found to lead poor performance. For remedy, a working model approach with rho=1 is proposed to provide valid inferences for testing the significance of covariate effects and estimating survival functions when rho is unknown. The proposed approaches are evaluated using simulation studies and illustrated in a large health screening data set.

email: yaob@email.sc.edu

A JOINT MODEL OF CANCER INCIDENCE, METASTASIS, AND MORTALITY

Qui Tran*, University of Michigan

Kelley M. Kidwell, University of Michigan

Alex Tsodikov, University of Michigan

Many diseases, especially cancer, are not static, but can be summarized as a series of events or stages (e.g. diagnosis, remission, recurrence, metastasis, death). We focus on cancer diagnosis, latent metastasis, and death as a sequence of cancer events. Most available methods to analyze multi-stage type of data ignore intermediate events and focus on the terminal one or consider (time to) multiple events as independent. Competing-risk or semi-competing-risk models often fall short of adequate description of the complex relationship between disease progression events driven by the shared progression process. A multi-stage model only looks at two stages at a time and fails to capture the effect of one stage on the time spent between other stages. Moreover, most models do not account for latent stages (e.g. onset of metastasis). Our semi-parametric joint model of diagnosis and latent metastasis events leading to cancer death uses nonparametric maximum likelihood to estimate covariate effects on the risks of intermediate events and death and the dependence between them. We illustrate the proposed method with Monte Carlo simulation and analysis of prostate cancer data from the SEER database.

email: *quitran@umich.edu*

PROPORTIONAL SUBDISTRIBUTION HAZARDS REGRESSION WITH INTERVAL-CENSORED COMPETING RISKS DATA

Yi Ren*, U.S. Food and Drug Administration

Chung-Chou Chang, University of Pittsburgh

Ruosha Li, University of Texas Health Science Center, Houston

WITHDRAWN

In survival analysis, the failure time of an event is not always exactly observed but interval-censored, where the event is only known to occur between two observation times. Most existing methods for interval-censored data only account for a single cause of failure. However, in many situations a subject may fail due to more than one type of event. Such data structures are called competing risks data. Competing events may include the occurrence of the event of interest. In the analysis of competing risks, the conventional methods may lead to nonsensical interpretation. With covariates, the proportional subdistribution hazards model is widely used to model the subdistribution of a particular event. For interval-censored competing risks data, however, estimation procedures based on the proportional subdistribution hazards model has not been investigated. In this dissertation, we propose estimation and inference procedures that account for both interval censoring and competing risks by adopting the modeling framework of the proportional subdistribution hazards model to examine the effects of covariates on the subdistribution. The proposed estimating equations effectively utilize the ordering of event time pairs. Simulation studies show that the proposed methods perform well under realistic scenarios. A lymphoma data set is used to illustrate the performance of the proposed method.

email: *yi.ren@fda.hhs.gov*

TUNING PARAMETER SELECTION IN COX PROPORTIONAL HAZARDS MODEL WITH A DIVERGING NUMBER OF PARAMETERS

Ai Ni*, Memorial Sloan Kettering Cancer Center

Jianwen Cai, University of North Carolina, Chapel Hill

Regularized variable selection methods are important tools for identifying the true model when the dimension of both the candidate models and the true model diverges with sample size. These methods involve one or more tuning parameters that control the complexity of the selected model. Therefore, the ability of the regularized variable selection methods to identify the true model critically depends on the correct choice of the tuning parameters. In this study we develop a consistent tuning parameter selection method for the Smoothly Clipped Absolute Deviation (SCAD) penalty under Cox's proportional hazards model with a diverging dimension. The selected tuning parameter minimizes the proposed Generalized Information Criteria (GIC), which is the negative log-partial likelihood penalized by a function of the model size. We prove that the selected tuning parameter leads to the true model with probability approaching one when sample size goes to infinity. Its finite sample performance is evaluated by simulations. It is applied to the Framingham Heart Study to identify the risk factors for the hazard of coronary heart disease.

email: *nia@mskcc.org*

PERMUTATION TEST FOR GENERAL DEPENDENT TRUNCATION

Sy Han Chiou*, Harvard School of Public Health

Jing Qian, University of Massachusetts, Amherst

Rebecca Betensky, Harvard School of Public Health

Quasi-independence is a common assumption for analyzing truncated survival data that are frequently encountered in biomedical science, astronomy, and social science. While the concept of censoring has been rigorously studied, many are not aware of the analytic issue that arise with delayed entry, or general truncation. Ignoring dependent truncation can severely bias estimation and inference. Current methods for testing for quasi-independent truncation are powerful for monotone alternatives, but not otherwise. Extending

methods in detecting highly non-monotone and even non-functional dependencies, we develop nonparametric tests that are powerful against non-monotone alternatives. We also describe and validate the use of an unconditional permutation method, which enables fixed risk-set-size permutation inference. The size and power of the propose d testing procedure are assessed in extensive simulation studies. An aging study in cognitive and functional decline is included to illustrate the usefulness of the method.

email: *schiou@hsph.harvard.edu*

SEMIPARAMETRIC MODELING AND ANALYSIS OF PAIRED LONGITUDINAL METHOD COMPARISON DATA

Lasitha N. Rathnayake*, University of Texas, Dallas

Pankaj K. Choudhary, University of Texas, Dallas

Method comparison studies are routinely conducted in biomedical disciplines to measure agreement between two methods of measuring a continuous response. Often, the measurements are taken over a period of time by both methods, giving rise to paired longitudinal data. We propose modeling the longitudinal profiles semiparametrically through penalized splines within the framework of mixed-effects models. The model allows the measurement errors to be correlated. Agreement between the methods is evaluated by performing inference on measures of agreement, such as concordance correlation coefficient and total deviation index, which are functions of parameters of the assumed model. Simulations show that this methodology performs well for moderately large number of subjects. The methodology is illustrated by analyzing a dataset of cholesterol measurements.

email: *lxr111030@utdallas.edu*

SEMIPARAMETRIC ESTIMATION OF THE ACCELERATED FAILURE TIME MODEL WITH PARTLY INTERVAL-CENSORED DATA

Fei Gao*, University of North Carolina, Chapel Hill

Donglin Zeng, University of North Carolina, Chapel Hill

Danyu Lin, University of North Carolina, Chapel Hill

Partly interval-censored (PIC) data arise when some failure times are exactly observed, while others are only known to lie within certain intervals. In this paper, we consider semiparametric efficient estimation of the accelerated failure time (AFT) model with PIC data. We first generalize the Buckely and James estimator for right-censored data to PIC data. Then, we develop a one-step estimator by deriving and estimating the efficient score for the regression parameters. We show that, under mild regularity conditions, the generalized Buckley-James estimator is consistent and asymptotically normal, and the one-step estimator achieves the semiparametric efficiency bound. We conduct extensive simulation studies to examine the performance of the proposed estimators in finite samples and apply our methods to data derived from a diabetes study.

email: *fgao@live.unc.edu*

79. STUDY DESIGN

VALUE-DRIVEN OPTIMIZATION OF STUDY DESIGN AND GO/NO GO DECISION AT POC STAGE: A PROGRAM LEVEL SIMULATION APPROACH

Masanori Ito*, Astellas Pharma Global Development Inc.

Nitin Patel, Cytel Inc.

Clinical trials at Proof-of-concept (PoC) stage are the critical milestone in the drug development. In general the available data of compound is limited at pre-PoC stage and therefore it is difficult to optimize the study design. It potentially includes complex trade-off problems for the cost and time versus the expected revenue and probability of success. For example, assume the situation that we choose either to set a couple of different dose groups of test drug or to set just one high dose group compared with placebo at the study designing stage. The latter is cheaper and faster than the former but the former provides richer data than the latter in terms of finding the minimum effective dose. Scenario simulations through the whole development program can support to make a good decision. Various dose response curves for efficacy and tolerability are assumed in the simulation to consider the trade-off between efficacy and safety. Expected net present value can be used to prioritize the scenarios. We do the simulations as consequence of clinical trials with various factors related to study design, Go/No Go decision criteria, time and cost. The simulation results are summarized and illustrated by some plots.

email: *masanori.ito@astellas.com*

A REVISIT TO TWO-WAY FACTORIAL ANOVA FOR UNBALANCED DATA

Tao Wang*, Medical College of Wisconsin

It has been known that the traditional ANOVA method can have biased estimates of variance components in two-way factorial ANOVA for unbalanced data. Typically, a general linear model (GLM) is used to estimate the variance components contributed by the main factors and their interactions in this case. However, the classical dummy variable based GLM often has the main effects and their higher-order interactions correlated even when the two treatment factors are independent. In this study, by adopting a mean-correction strategy, we propose a revised GLM to dissect the confounding between the main effects and their interactions. We

show that this revised GLM can provide an orthogonal partition on the variance components when the two treatment factors are independently assigned. It can be fitted using the standard least square approach. It also allows us to conveniently test for the existence of variance components via hypothesis testing of the regression coefficients for the main effects and their interactions.

email: *taowang@mcw.edu*

EXPOSURE ENRICHED CASE-CONTROL (EECC) DESIGN FOR THE ASSESSMENT OF GENE-ENVIRONMENT INTERACTION

Md Hamidul Huque*, University of Technology Sydney, Australia

Raymond J. Carroll, Texas A&M University

Nancy Diao, Harvard School of Public Health

David C. Christiani, Harvard School of Public Health

Louise M. Ryan, University of Technology Sydney, Australia

Genetic susceptibility and environmental exposure both play an important role in the aetiology of many diseases. Case-control studies are often the first choice to explore the joint influence of genetic and environmental factors on the risk of developing a rare disease. In practice, however, exposure distributions may be highly skewed, in which case, power to detect rare disease effects can be limited, especially when susceptibility genes are rare. We propose a variant of the classical case-control study where not only cases, but also high (or low) exposed individuals are oversampled, depending on the skewness of the exposure distribution. Of course, a traditional logistic regression model is no longer valid and therefore results in biased estimation. We show that the addition of a simple covariate to the regression model removes this bias. We show that our proposed exposure enriched case-control (EECC) design provides reliable estimates of main and interaction effects of interest. We also show that judicious over-sampling of high/low exposed individuals can boost study power. We illustrate our results using an example involving arsenic exposure and detoxification genes in Bangladesh. Our study reveals that statistical power for the detection of gene environment interactions can be enhanced with a simple extension of the classical case-control design.

email: *MdHamidul.Huque@student.uts.edu.au*

SAMPLE SIZE CALCULATIONS FOR STRATIFIED MICRO-RANDOMIZED TRIALS IN mHEALTH

Walter Dempsey*, University of Michigan

Peng Liao, University of Michigan

Susan Murphy, University of Michigan

Technological advancements in the field of mobile devices and wearable sensors have helped overcome obstacles in the delivery of care, making treatment available anytime and anywhere. These treatments are often designed to have near-term impact on individuals; yet it is often unclear whether treatment effects occur and if so, when and in which context, the treatments are most effective. Scientific excitement in the potential of mobile interventions has led to development far outpacing the design of corresponding statistical methods. With this paper, we introduce the "Stratified Micro-randomized Trial", in which each individual is randomized among treatments 100's or 1000's of times. Furthermore the randomization probabilities depend on a time-varying covariate, which may be an outcome of past treatment. We develop two approaches to determining sample size. In the first approach the primary hypothesis concerns testing for a marginal treatment effect, marginal over the time-varying covariate used in stratification. The second approach concerns testing for the treatment effect conditional on the time-varying covariate. We address the trade-offs between these two approaches and provide associated sample size calculators. This work is motivated by a mobile health smoking cessation study in which randomization probabilities should depend on a binary time-varying stress classification.

email: *wdem@umich.edu*

COMPOUND CRITERIA FOR CONSTRUCTING EFFICIENT AND FLEXIBLE DESIGNS

Luzia A. Trinca*, Universidade Estadual Paulista, Brasil (UNESP)

Standard optimal design criteria were developed under the assumption that an independent error variance estimate would be available or that the model considered prior to experimentation would be the correct one. However, in practice, prior error variance estimate is hardly available and particularly for quantitative factors, the model assumed is just an approximation to the true relation between the reponse and treatment effects. Recently some modified optimality criteria were defined, which correctly reflect the utility of designs with respect to some common types of inference, and approaches based on multiple objectives were introduced in order to construct more robust designs. By using compound criteria that incorporate several desired design properties it is possible to produce designs that are likely to be more relevant to many practical situations. Two examples illustrate the advantges of the methods in relation to standard approaches. The first example deals with a tissue culture experiment with many factors but restricted amount of explants while the second considers a water metal removal experiment with randomization restrictions.

email: *ltrinca@ibb.unesp.br*

SeqDesign: A FRAMEWORK FOR RNA-Seq GENOME-WIDE POWER CALCULATION AND EXPERIMENTAL DESIGN ISSUES

Chien-Wei Lin*, University of Pittsburgh

Serena G. Liao, University of Pittsburgh

George C. Tseng, University of Pittsburgh

Next Generation Sequencing (NGS) technology is emerging as an appealing tool in characterizing genomic profiles at individual level. In particular, RNA-seq is becoming a standard tool for global transcriptomic monitoring. Although the experimental cost continues to drop rapidly, the high sequencing expense and bioinformatic complexity will continue to be an obstacle for many biomedical projects. Modelling of NGS data not only involves sample size and genome-wide (multiple comparison) power inference, but also includes consideration of sequencing depth and count data property. Consequently, given total budget and pre-specified cost parameters, the experimental design issue in RNA-seq is conceptually a much more complex optimization problem than the traditional univariate hypothesis testing and one-dimensional sample size calculation scenario. In this paper, we propose a "SeqDesign" statistical framework to utilize pilot data for power calculation and experimental design of RNA-seq experiments. Our approach is based on mixture model fitting of p-value distribution from pilot data and a parametric bootstrap procedure based on approximated Wald test statistics to infer genome-wide power in targeted sample size and sequencing depths. Realistic experimental design tasks are illustrated. We perform simulation and one real application to illustrate its performance compared with existing methods. An R package "SeqDesign" is publicly available.

email: *masaki396@gmail.com*

STATISTICAL CONSIDERATIONS IN DESIGNING PRECISION STUDY FOR OPTICAL COHERENCE TOMOGRAPHY DEVICE

Haiwen Shi*, U.S. Food and Drug Administration

Optical Coherence Tomography (OCT) is a medical imaging technology invented in 1991. Since its invention, the OCT has been applied for imaging of eyes and has had largest impact in ophthalmology. OCT has been used for diagnosis and monitoring of retinal diseases such as glaucoma. It is critical for OCT to accurately measure and monitor the thickness of eye anatomy such as retina and retinal nerve fiber layer, which are associated with progression of some ocular disease. Hence, the OCT needs to have good precision in the measurement of the thickness. In this talk, I will discuss some statistical considerations in designing a good precision study for OCT. A typical precision study for OCT uses either nested or crossed design, which means different subjects or same subjects are scanned under each OCT + operator configuration, respectively. The pros and cons of nested and crossed design will be discussed. The discussion will be both in analytical and through simulation. By using simulation, the variance components of the OCT + operator configurations and the repeatability and reproducibility estimates for two designs are compared. In addition, I will discuss other related issues such as the alternative way to estimate repeatability and reproducibility.

email: *haiwen.shi@fda.hhs.gov*

80. PRESIDENTIAL INVITED ADDRESS

BIOSTATISTICS, BIOMEDICAL INFORMATICS, AND HEALTH DATA SCIENCE: RESEARCH AND TRAINING

Xihong Lin, Ph.D., Chair and Henry Pickering Walcott Professor, Department of Biostatistics, Harvard University

Biostatistics has played a pivotal role in both the development and success of basic science, public health, and medical research by developing statistical methods for study design and data analysis. Massive 'ome data, including genome, exposome, and phenome data, are becoming available at an increasing rate with no apparent end in sight. Examples include Whole Genome Sequencing data, large-scale remote-sensing satellite air pollution data, digital phenotyping data, and Electronic Medical Records. The emerging field of Health Data Science (HDS) presents biostatisticians with many research and training opportunities and challenges. It has propelled us to rethink our identity and niche and how we can properly position ourselves as a leader in HDS, especially in promoting and advancing statistical inference in health data science research and training. Success will both for biostatistics and for much of health and biomedical science that we effectively position ourselves together with bio- and medical informaticians, as leading health data scientists. There are countless of examples where the volume of available data requires new, scalable statistical methods and demand an investment in statistical research. These include signal detection, network analysis, integrated analysis of different types and sources of data, and incorporation of domain knowledge in health data science method development. Especially critical is training the next generation of health data scientists, which include not only providing broader training of health and biomedical researchers in sound statistical inference, but also that integrate computer and information science and machine learning into established biostatistical curriculum. Such enhanced training could include both didactic and EdX courses, but will require a careful balance of depth and breadth across areas. In this talk, I discuss some of the challenges and opportunities, and illustrate them using statistical genetics and genomics as examples.

e-mail: *xlin@hsph.harvard.edu*

81. NEW DEVELOPMENTS FOR INDIVIDUALIZED MEDICAL DECISION MAKING IN REAL WORLD SETTINGS

DYNAMIC SYSTEMS FOR IDENTIFYING BIOMARKERS PREDICTING LANDMARKS OF DISEASE DEGENERATION

Yuanjia Wang*, Columbia University

Precise modeling of disease progression in neurodegenerative disorders may enable early intervention before clinical manifestation of disease, which is crucial since early intervention in premanifest subjects is expected to be more effective. Neuroimaging biomarkers are indicative of the underlying disease pathology and may be used to predict disease at the preclinical stage. As observed in many empirical studies, longitudinal measurements of clinical outcomes, such as motor or cognitive symptoms, often present nonlinear sigmoid shapes over time, where inflection points in the trajectory mark a critical time in disease progression. Therefore, to identify neuroimaging biomarkers predicting the disease progression, we propose a nonlinear mixed effects model based on a sigmoid function to predict longitudinal clinical outcomes, and associate a high-dimension linear predictor of neuroimaging biomarkers with subject-specific inflection points. Variable selection is incorporated in the algorithm in order to identify important biomarkers of disease progression and reduce prediction variability. We discuss implications of our models on phase 2 clinical trials.

email: *yw2016@columbia.edu*

MODEL VALIDATION AND SELECTION IN G-ESTIMATION OF DYNAMIC TREATMENT REGIMES

Erica E. M. Moodie*, McGill University

Michael P. Wallace, McGill University

David A. Stephens, McGill University

Dynamic treatment regimes (DTRs) recommend treatments based on evolving subject-level data, and therefore the identification of an optimal DTR is a key goal for precision medicine. Estimation of such regimes using semi-parametric approaches which are doubly robust have gained popularity, as they afford some degree of protection in settings where not all models are easily specified. However, in practice it can be difficult to assess whether models have been correctly-specified outside of the special setting of sequentially randomized trials. In this talk, I will demonstrate some model validation and selection tools that have been especially adapted for use in doubly robust g-estimation of optimal DTRs.

email: *erica.moodie@mcgill.ca*

LEARNING OPTIMAL PERSONALIZED TREATMENT RULES IN BENEFIT-RISK ANALYSIS

Yuanjia Wang, Columbia University

Haoda Fu, Eli Lilly and Company

Donglin Zeng*, University of North Carolina, Chapel Hill

Personalized medicine has become a more and more important issue in the new era of medical product development. While an optimal treatment for a patient often aims to maximizes clinical benefit, it may also lead to high concern of safety and possible adverse events. Therefore, benefit and risk should be considered simultaneously when estimating the optimal personalized treatment rules. We propose a new learning approach to identify personalized optimal treatment strategy that maximizes clinical benefit under a constraint for the risk. We extend existing regression-based learning approaches (Q-learning) and weighted learning approaches (O-learning) to estimate the optimal rules in this benefit-risk context. The algorithms, simulations, and theoretical properties of the proposed method will be presented. Finally, we apply our approach to a randomized trial of type 2 diabetes to guide optimal administration of the first line insulin treatments based on individual patient characteristics while controlling for the number of hypoglycemia events.

email: *dzeng@email.unc.edu*

BUILDING PERSONALIZED TREATMENT STRATEGY WITH BINARY OUTCOMES

Min Qian*, Columbia University

Eric Laber, North Carolina State University

A treatment policy is a sequence of decision rules that specify how the dosage and/or type of treatment should be adjusted through time in response to an individual's changing needs. Q-learning, which involves an iterative two-step procedure that first uses regression to model the conditional mean outcome at each stage, and second, derives the estimated policy by maximizing the estimated conditional mean functions, is often used on data from SMART studies to develop the optimal treatment policy. We propose to generalize Q-learning to the case of binary outcome with data from a partial SMART study, where only a proportion of patients went through the full course of the trial. The method will be illustrated using data from a web-based smoking cessation study.

email: *mq2158@columbia.edu*

82. EMERGING ISSUES IN CLINICAL TRIALS WITH TIME-TO-EVENT DATA IN THE PRESENCE OF COMPETING RISKS

A BAYESIAN CURE RATE FRAILTY MODEL FOR SURVIVAL DATA IN PRESENCE OF SEMI-COMPETING AND COMPETING RISKS

Mario de Castro, Universidade de Sao Paulo, Brasil

Ming-Hui Chen*, University of Connecticut

Anthony V. D'Amico, Harvard University and Brigham and Women's Hospital

Semi-competing risks data include the time to a nonterminating event and the time to a terminating event while competing risks data include the time to more than one terminating events. Our study is motivated from a prostate cancer study, which has one nonterminating event and two terminating events with both semi-competing risks and competing risks present. Due to the complication of two non-informative censoring times for the nonterminating event and the terminating events, the existing semi-competing risks models may not be identifiable. In this paper, we propose a new cure rate frailty model for this type of survival data. The proposed model is not only identifiable but also theoretically and computationally attractive. In addition, the proposed model can easily accommodate non-informative right-censoring times for the nonterminating and terminating events. The properties of the proposed model are examined in detail and an efficient Markov chain Monte Carlo sampling algorithm is also developed. The proposed methodology is further assessed using simulation as well as an analysis of the real data from a prostate cancer study.

e-mail: *ming-hui.chen@uconn.edu*

TREATMENT EFFECT ESTIMATE AND MODEL DIAGNOSTICS WITH TIME-VARYING TREATMENT SWITCHING

Qingxia Chen*, Vanderbilt University

Fan Zhang, University of Connecticut

Ming-Hui Chen, University of Connecticut

Xiuyu Julie Cong, Boehringer Ingelheim Pharmaceuticals, Inc.
Treatment switching frequently occurs in clinical trials due to ethical reasons. Intent-to-treat analysis without adjusting for switching yields biased and inefficient estimates of the treatment effects. In this paper, we propose a class of semiparametric semi-competing risks transition survival models to accommodate time-varying switches. Theoretical properties of the proposed model are examined. An efficient expectation-maximization algorithm is derived and implemented in existing software for obtaining the maximum likelihood estimates. Model diagnostic tools are developed for the proposed models. Simulation studies are conducted to demonstrate the validity of the model. The proposed method is further applied to data from a clinical trial with patients having recurrent or metastatic squamous-cell carcinoma of the head and neck.

email: *cindy.chen@vanderbilt.edu*

REGRESSION ANALYSIS FOR CUMULATIVE INCIDENCE FUNCTION UNDER TWO-STAGE RANDOMIZATION

Idil Yavuz, Dokuz Eylul University, Turkey

Ling-Wan Chen, University of Pittsburgh

Yu Cheng*, University of Pittsburgh

Abdus Wahed, University of Pittsburgh

In this talk we focus on regression analysis of multiple event data from a two-stage randomization trial. Even though there is extensive research on the regression problem for dynamic treatment regimes, few research has considered competing risks outcomes which commonly occur in practice with multiple endpoints. We will focus on modeling the cumulative incidence function (CIF) of a cause-specific outcome, and extend some commonly used regression models in the competing risk literature, such as the multi-state, Fine and Gray, and Scheike et al. regression models, to the two-stage randomization setting. Through the augmentation of the data, the proposed models can be implemented in R using the existing packages. We show the improvement our methods provide by simulation.

email: *yucheng@pitt.edu*

PENALIZED VARIABLE SELECTION IN COMPETING RISKS REGRESSION

Zhixuan Fu, Yale University

Chirag Parikh, Yale University

Bingqing Zhou*, Novartis and Yale University

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 multiple mutually exclusive events, known as competing risks. The proportional subdistribution hazard (PSH) model proposed by Fine and Gray (1999) has become a popular semi-parametric regression model for time-to-event data with competing risks. In this paper, we propose a general penalized variable selection strategy for the PSH model that simultaneously handles variable selection and parameter estimation. We rigorously establish the asymptotic properties of the proposed penalized estimators and modify the coordinate descent algorithm for implementation. Extensions involving stratification and group variable selection are also addressed. Simulation studies are conducted to

demonstrate the good performance of the proposed method. Data from a single arm oncology clinical trial serve for an illustration.

email: *bingqing.zhou@yale.edu*

83. NEW DEVELOPMENTS OF STATISTICAL METHODS FOR FAMILY-BASED SEQUENCING STUDIES

RARE-VARIANT ASSOCIATION TESTING OF COMPLEX DISEASE IN PEDIGREES USING IDENTITY-BY-DESCENT INFORMATION

Michael P. Epstein*, Emory University School of Medicine

Glen A. Satten, Centers for Disease Control and Prevention

While many rare-variant association tests exist for case-control designs, far fewer tools exist for studying disease in affected pedigrees. This is unfortunate since affected pedigrees possess many attractive features for rare-variant analysis that case-control studies lack. As a result, there is increased interest in sequencing familial samples, particularly those collected from past linkage projects. We recently published an approach for rare-variant testing in affected sibships [AJHG 96: 543] based on the idea that rare susceptibility variants should be found more on regions shared identical by descent (IBD) by affected siblings than on regions not shared IBD. The strategy is more powerful than analogous case-control association testing and is also robust to population stratification. Here, we expand our framework further to allow more flexible analyses of family-based NGS studies of disease. We show how to extend the IBD framework to allow general rare-variant analysis of affected pedigrees of arbitrary structure such that any pedigree-based NGS study can use our methods. We also develop a novel two-stage screening and validation procedure for rare-variant analysis that has the benefit of using the same set of affected pedigrees for both stages. The screening test compares rare-variant burden of affected relatives from the pedigrees to the burden of external controls, while the independent validation test compares rare-variant burden from the same affected relatives to their IBD sharing in the top first-stage signals. This strategy provides valid and powerful replication of initial rare-variant findings, without requiring additional sample collection. We illustrate our approaches using existing NGS studies of complex traits. This is joint work with Dr. Glen Satten (Centers for Disease Control).

email: *mepstein@genetics.emory.edu*

SEARCHING RARE VARIANTS UNDER COMPLEX TRAITS LEVERAGING ON LINKAGE EVIDENCE

Xiaofeng Zhu*, Case Western Reserve University

Many statistical methods for analyzing rare variant association have been developed but most of them focus on unrelated samples. Family based linkage analysis had been extensively studied before genome wide association studies (GWASs) and has recently been much ignored. Family based linkage has many advantages, such that it is immune to genetic heterogeneity, therefore, could be potentially powerful in detecting rare variants underlying a complex trait. In addition, many genetic variants detected by GWASs cumulatively account for limited phenotypic variation and unable to explain linkage evidence reported in literature. Here we introduce a computation efficient rare variant association approach leveraging on linkage evidence. We will illustrate the statistical properties using simulations and real data as well. We will also compare the proposed method with existing methods such as famSKAT. We will demonstrate that family data are useful in searching for very rare variants underlying complex traits.

email: *xiaofeng.zhu@case.edu*

GENE-BASED ASSOCIATION TESTING OF DICHOTOMOUS TRAITS USING GENERALIZED FUNCTIONAL LINEAR MIXED MODELS FOR FAMILY DATA

Yingda Jiang*, University of Pittsburgh

Chi-Yang Chiu, 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

Qi Yan, Children's Hospital of Pittsburgh of the University of Pittsburgh Medical Center

Wei Chen, Children's Hospital of Pittsburgh of the University of Pittsburgh Medical Center

Michael B. Gorin, University of California, Los Angeles

Yvette P. Conley, University of Pittsburgh

Daniel E. Weeks, University of Pittsburgh

Gene-based association testing requires aggregating the variant information in the gene into a single measure. As genotyping data can be viewed as a realization of a stochastic process varying along the chromosome, the genetic information can be summarized using functional data analysis where discrete genotypes are fitted by a continuous curve by using a collection of smooth basis functions. Gene-based association tests for dichotomous traits and unrelated samples have been developed using generalized functional linear models (FLMs). In most situations, these tests have higher power than well-known kernel-based methods (e.g., SKAT-O). Here we extend this approach to family-based data using the GLOGS (genome-wide logistic mixed model/score test) approach of Stanhope and Abney. This involves parallel computations to integrate out a multidimensional polygenic effect. Simulation results indicate that our new statistics are better than other similar statistics (famSKAT

or F-SKAT), but not better than the retrospective kernel and burden statistics of Schaid and colleagues. We also embed FLM-smoothed genotypes into the retrospective statistics, improving the power of the kernel-based approach. We apply these statistics to an age-related macular degeneration (AMD) family data set, where, as expected, we observe strong association between AMD and CFH and ARMS2, two known AMD susceptibility genes.

email: *weeks@pitt.edu*

A BAYESIAN FRAMEWORK FOR DE NOVO MUTATION CALLING IN FAMILY SEQUENCING DATA

Qiang Wei, Vanderbilt University

Rui Chen, Vanderbilt University

Xue Zhong, Vanderbilt University

Yongzhuang Liu, Harbin Institute of Technology

Xiaowei Zhan, University of Texas Southwestern

Wei Chen, University of Pittsburgh

Bingshan Li*, Vanderbilt University

Spontaneous (de novo) mutations play an important role in the disease etiology of a range of complex diseases. Identifying de novo mutations (DNMs) in sporadic cases provides an effective strategy to find genes or genomic regions implicated in the genetics of disease. High-throughput next-generation sequencing enables genome- or exome-wide detection of DNMs by sequencing probands and their parents and siblings. Such approaches have been employed to search for genes implicated in various diseases including autism, schizophrenia and epilepsy. For sequencing studies the traditional approach to DNM calling is to call individual genotypes separately in a pedigree and then compare the offspring and parental genotypes to identify DNMs. This naive approach is inevitably ineffective and often generates many false positive DNM calls due to sequencing error and alignment artifacts. Here we describe a novel Bayesian framework for DNM calling, which jointly models sequencing data in a pedigree and naturally overcomes some limitations inherent in other calling methods. The calling accuracy is further improved when the identity-by-descent allele sharing among siblings is modeled in the framework. We illustrate the performance of the new framework compared to other state-of-the-art methods on both simulated and real datasets.

email: *bingshan.li@vanderbilt.edu*

84. NEW DEVELOPMENTS OF QUANTILE REGRESSION FOR COMPLEX DATA ANALYSIS: THEORIES AND APPLICATIONS

PARTIALLY LINEAR ADDITIVE QUANTILE REGRESSION IN ULTRA-HIGH DIMENSION

Ben Sherwood, John Hopkins University

Lan Wang*, University of Minnesota

We consider a flexible semiparametric quantile regression model for analyzing high dimensional heterogeneous data. This model has several appealing features: (1) By considering different conditional quantiles, we may obtain a more complete picture of the conditional distribution of a response variable given high dimensional covariates. (2) The sparsity level is allowed to be different at different quantile levels. (3) The partially linear additive structure accommodates nonlinearity and circumvents the curse of dimensionality. (4) It is naturally robust to heavy-tailed distributions. In this paper, we approximate the nonlinear components using B-spline basis functions. We first study estimation under this model when the nonzero components are known in advance and the number of covariates in the linear part diverges. We then investigate a non-convex penalized estimator for simultaneous variable selection and estimation. We derive its oracle property for a general class of non-convex penalty functions in the presence of ultra-high dimensional covariates under relaxed conditions. To tackle the challenges of nonsmooth loss function, non-convex penalty function and the presence of nonlinear components, we combine a recently developed convex-differencing method with modern empirical process techniques. Monte Carlo simulations and an application to a microarray study demonstrate the effectiveness of the proposed method. We also discuss how the method for a single quantile of interest can be extended to simultaneous variable selection and estimation at multiple quantiles.

e-mail: *wangx346@umn.edu*

MODEL SELECTION FOR QUANTILE REGRESSION WITH VARYING COVARIATE EFFECTS

Qi Zheng, University of Louisville

Limin Peng*, Emory University

Varying covariate effects often manifest meaningful heterogeneity in covariate-response associations. Contemporaneously examining a spectrum of quantiles under quantile regression allows for identifying variables with either partial or full effects. Under this motivation, we study a general adaptively weighted LASSO penalization strategy for varying-coefficient quantile regression, where a continuum of quantile index is considered. For the finite p case, we

establish the desirable oracle properties. Furthermore, we formally investigate a BIC-type uniform tuning parameter selector and show that it can ensure consistent model selection. Our numerical studies confirm the theoretical findings and illustrate an application of the new method.

e-mail: *lpeng@sph.emory.edu*

REGULARIZED QUANTILE REGRESSION UNDER HETEROGENEOUS SPARSITY WITH APPLICATION TO QUANTITATIVE GENETIC TRAITS

Chad He*, Fred Hutchinson Cancer Research Center

Linglong Kong, University of Alberta

Yanhua Wang, Beijing Institute of Technology

Sijian Wang, University of Wisconsin, Madison

Timothy Chan, Memorial Sloan-Kettering Cancer Center

Eric Holland, Fred Hutchinson Cancer Research Center

Genetic studies often involve quantitative traits. Identifying genetic features that influence quantitative traits can help to uncover the etiology of diseases. Quantile regression method considers the conditional quantiles of the response variable, and is able to characterize the underlying regression structure in a more comprehensive manner. On the other hand, genetic studies often involve high-dimensional genomic features, and the underlying regression structure may be heterogeneous in terms of both effect sizes and sparsity. We introduce a regularized quantile regression method that is able to account for the potential genetic heterogeneity. We investigate the theoretical property of the proposed method, and examine its performance through a series of simulation studies. A real dataset is analyzed to demonstrate the application of the proposed method.

e-mail: *qhe@fhcrc.org*

SOME ASPECTS OF REGULARIZATION IN QUANTILE REGRESSION

Ivan Mizera*, University of Alberta

Aspects of various prescriptions involving use of penalties in quantile regression will be reviewed and discussed. These include constrained (on penalty and loss function) as well as unconstrained formulations, and various difference and differential operators. The examples include regularization involving total variation, various smoothing penalties, and various sparsity-promoting schemes. The connections stemming from convex duality are explored as well.

e-mail: *imizera@yahoo.com*

85. CURRENT DEVELOPMENTS AND ISSUES FOR META-ANALYSIS

SOME RECENT THEORETICAL RESULTS ON META-ANALYSIS

Danyu Lin*, University of North Carolina, Chapel Hill

Motivated by applications to genetic association studies, we have investigated some theoretical issues in meta-analysis in recent years. In two companion Biometrika papers (Lin and Zeng, 2010; Zeng and Lin, 2015), we showed that meta-analysis based on summary statistics is asymptotically equivalent to joint analysis of individual-participant data under fixed-effects models and is at least as efficient as the latter under random-effects models. In this talk, I will describe some newer theoretical results. First, I will present a new class of test statistics for simultaneous inference on the mean and heterogeneity of effect sizes and show that the use of summary statistics is equivalent to the use of individual-participant data for this type of tests. Second, I will show that it is possible to test the conditional effect of one variable on the outcome given another (correlated) variable using only summary statistics, such that one can perform conditional meta-analysis without fitting the conditional models. I will illustrate the theoretical results with empirical data from genetic association studies.

email: *lin@bios.unc.edu*

ADAPTIVELY WEIGHTED META-ANALYSIS IN -OMICS APPLICATIONS

Zhiguang Huo, University of Pittsburgh

Yongseok Park, University of Pittsburgh

George Tseng*, University of Pittsburgh

In this talk, I will present an adaptively weighted Fisher's method that was developed to account for gene-specific heterogeneous differential expression (DE) signal across studies in transcriptomic meta-analysis. We will show that the method is asymptotically Bahadur optimal and the adaptive weights provide improved interpretation and further biological investigation in omics applications. A fast algorithm has been developed for accurate p-value calculation that is useful for multiple comparison of thousands of genes (or millions of SNPs). Bootstrap technique is used to assess a confidence score for the weights. Simulations and applications of prostate cancer, lung cancer and mouse metabolism data will be presented.

email: *ctseng@pitt.edu*

NETWORK META-ANALYSIS FOR DIAGNOSTIC ACCURACY

Thomas Trikalinos*, Brown University

Wei Cheng, Brown University

Constantine Gatsonis, Brown University

Christopher Schmid, Brown University

Constructing a network of diagnostic test accuracy studies in order to compare multiple tests is more complex than doing so for studies of treatment efficacy. Synthesizing diagnostic accuracy studies may focus on summarizing the diagnostic performance of each single test rather than the pairwise contrast. It shall include information from eligible subjects with single-, paired- and triplet-test studies for each test, and because the TPF (true positive fraction, which equals sensitivity) and FPF (false positive fraction, which equals 1-specificity) of test(s) in a diagnostic accuracy study are correlated. We propose a joint modeling framework for networks of diagnostic accuracy studies with mixed study-types (single-, paired- and triplet-test studies). The model assumes that true and false positive counts follow Binomial distributions independently among diseased and nondiseased individuals. The underlying true and false positive fractions for each test are decomposed on the logit scale into components that represent their overall average across study-types for each test, study-type specific effects to reflect inconsistency, and within-study-type random effects to account for heterogeneity (with the concepts of inconsistency and heterogeneity in agreement with Higgins et al. 2012). The method is applied to a network of studies testing the utility of multiple biomarkers obtained by second-trimester prenatal ultrasounds for the detection of trisomy 21 (Downâ€™s syndrome) in fetuses.

email: *thomas_trikalinos@brown.edu*

A NOVEL METHOD FOR CORRECTING PUBLICATION BIAS IN MULTIVARIATE META-ANALYSIS

Yong Chen*, University of Pennsylvania

Chuan Hong, University of Texas Health Science Center, Houston

Haitao Chu, University of Minnesota

Publication bias occurs when the publication of research results depends not only on the quality of the research but also on its nature and direction. A consequence is that published studies may not be truly representative of all valid studies undertaken, and the corresponding bias may threaten the validity of systematic reviews. Multivariate meta-analysis has recently received increasing attention for its ability reducing potential bias and improving statistical efficiency by borrowing information across outcomes. However, detecting and accounting for publication bias are more challenging in multivariate meta-analysis setting. In this talk, I will describe a novel multivariate method for accounting for publication bias. The proposed method is not only powerful in detecting publication bias, but also can effectively correct publication bias. The proposed method is validated through simulation studies and is illustrated through case studies (Joint work with C Hong and HT Chu).

email: *ychen123@mail.med.upenn.edu*

86. SURVIVAL PREDICTION MODELS FOR MEDICAL DECISION MAKING

EVALUATION OF BIOMARKERS FOR PREDICTION OF CLINICAL EVENTS: CONNECTION TO INFORMATION THEORY

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 predictive potential or incremental value are needed. We detail new non-parametric methods that connect partial likelihood and information theory criteria. Our methods can be used to detail biomarker performance over time, and provide a single information summary measure that can rank and/or compare markers.

email: *heagerty@uw.edu*

DYNAMIC PREDICTION OF TIME-TO-EVENT DISTRIBUTIONS

Xuelin Huang*, University of Texas MD Anderson Cancer Center

Fangrong Yan, China Pharmaceutical University and University of Texas MD Anderson Cancer Center

Jing Ning, University of Texas MD Anderson Cancer Center

Ziding Feng, University of Texas MD Anderson Cancer Center

Dynamic prediction is to use longitudinal biomarkers for real-time prediction of individual prognosis. This is critical for patients with non-curable disease such as cancer. Their longitudinal biomarker values fluctuate over time, and the changing patterns vary greatly across patients. These variations make it a difficult task to model the longitudinal biomarker values over time. In this talk, we propose two approaches. One is to model biomarker trajectories by functional data analysis techniques, and link the features of trajectory functions to the risk of disease progression. The second approach is to avoid modeling the changing patterns of longitudinal biomarkers, but assume that their effects on disease recurrence risks are smooth functions of the prediction time. The proposed methods are evaluated by simulation studies, and applied to make dynamic predictions for patients with chronic myeloid leukemia at any time following their treatment with tyrosine kinase inhibitors, using their longitudinally measured BCR-ABL gene expression levels to predict their risk of disease progression.

email: *xlhuang@mdanderson.org*

ROBUST LEARNING OF OPTIMAL TREATMENT REGIMES FOR SURVIVAL ENDPOINTS

Runchao Jiang, Facebook

Wenbin Lu*, North Carolina State University

Rui Song, North Carolina State University
Michael Hudgens, University of North Carolina, Chapel Hill
Sonia Naprvavnik, University of North Carolina, Chapel Hill

Individualized treatment regimes have the potential to improve clinical outcomes of interest. When the primary outcome is survival time, the literature discusses how to identify the optimal regime to maximize the survival probability at a particular time point t. However, t-year survival probability may not be a good choice, since it is unable to balance the short-term and long-term benefits. In this paper, we proposed a doubly robust approach to identify the optimal regime, which optimizes some comprehensive functions of the survival curve. Two commonly used examples are the restricted mean survival time and the median survival time. We investigated the empirical performance of the proposed methods via simulation studies and established their asymptotic properties. We also applied the proposed methods to a UNC AIDS data, and showed that the proposed methods significantly improve the restricted mean time of the initial treatment duration.

email: *lu@stat.ncsu.edu*

AN ANALYTICAL FRAMEWORK FOR BUILDING AND EVALUATING LANDMARK MODELS FOR DYNAMIC PREDICTION OF SURVIVAL USING LONGITUDINAL DATA

Liang Li*, University of Texas MD Anderson Cancer Center
Sheng Luo, University of Texas Health Science Center, Houston
Bo Hu, Cleveland Clinic
Tom Greene, University of Utah

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 as well as other time-dependent and time-independent information about the patient. The prediction is dynamic in the sense that it can be made at any time during the follow-up, adapting to the changing at-risk population and incorporating the most recent longitudinal data. One approach is to build a joint model of longitudinal predictor variables and time to the clinical event, and draw predictions from the posterior distribution of the time to event conditional on longitudinal history. Another approach is to use the landmark model, which is a system of prediction models that evolve with the follow-up time. We review the pros and cons of the two approaches, and present a general analytical framework using the landmark approach. The proposed framework allows the measurement times of longitudinal data to be irregularly spaced and differ between subjects. We propose a unified kernel weighting approach for estimating the model parameters, calculating predicted probabilities, and evaluating prediction accuracy through double time-dependent Receiver Operating Characteristics (ROC) curves. We illustrate the proposed analytical framework using the African American Study of Kidney Disease and Hypertension (AASK) to develop a landmark model for dynamic prediction of end stage renal diseases or death among patients with chronic kidney disease.

email: *LLi15@mdanderson.org*

87. STATISTICAL MACHINE LEARNING FOR BIG-BIO-DATA

ESTIMATION OF DIRECTED ACYCLIC GRAPHS USING BIC UNDER PATH RESTRICTIONS

George Michailidis*, University of Florida

Directed Acyclic Graphs (DAGs) are frequently used to represent complex causal processes in many application areas. The edges of a DAG generally encode dependence relationships and give a factorization of a joint likelihood, often multivariate Gaussian, for the nodal variables. The edges, or, depending on the setting, their undirected counterparts, can be estimated from observations of these variables using conditional independence tests as in the PC algorithm or penalized-likelihood approaches such as the lasso or BIC. These methods have been well studied in general, but limited attention has been given to estimation in settings where the space of allowable DAG structures is constrained by prior information. For instance, in applications such as functional genomics a limited set of perturbation screens may have previously revealed that certain paths are present and others absent. In this work we define incomplete partial orders as a formal way for describing such path-based restrictions on DAGs. Incomplete partial orders generalize the important special case of a known linear order. We then develop a stochastic optimization algorithm for obtaining estimates using BIC over the restricted space and explore its performance in examples. In addition, we present a simple visualization for comparing the information content in different incomplete partial orders.

email: *gmichail@umich.edu*

SPATIALLY RELATING DEVELOPMENTAL TRANSCRIPTION FACTORS USING DROSOPHILA EMBRYONIC GENE EXPRESSION IMAGES

Karl Kumbier*, University of California, Berkeley
Siqi Wu, University of California, Berkeley
Antony Joseph, University of California, Berkeley
Ann Hammonds, Lawrence Berkeley National Laboratory
William Fisher, Lawrence Berkeley National Laboratory
Richard Weiszmann, Lawrence Berkeley National Laboratory
Sue Celniker, Lawrence Berkeley National Laboratory

Bin Yu, University of California, Berkeley

Erwin Frise, Lawrence Berkeley National Laboratory

Spatially defined gene interactions have long been known to take part in developmental processes. The recent abundance of spatial gene expression data is opening up new opportunities to understand gene-gene interactions that behave uniquely across different regions of developing embryos. Given the scale and complexity of such data, interpretable models are necessary to guide inquiry and guard against false discoveries. Using non-negative matrix factorization, we analyze spatial gene expression patterns from a large microscopy dataset of Drosophila melanogaster embryos, providing low-dimensional, biologically meaningful representations. Based on our "principal patterns" of gene expression, we construct spatially local correlation networks that correspond well to the gap gene network in early stage embryos. We also discuss our work to extend these results to late stage Drosophila, where morphology across embryos is highly inconsistent. To handle these inconsistencies, we have developed an organ classification and registration model that modifies state of the art computer vision algorithms to produce mid-level image features well suited to bioimaging tasks. Combined with our principal patterns, this pipeline is a promising step towards a late stage analogue to the Drosophila fate map.

email: *kkumbier@berkeley.edu*

ESTIMATING FALSE INCLUSION RATES IN PENALIZED REGRESSION MODELS

Patrick Breheny*, University of Iowa

Penalized regression methods are an attractive tool for feature selection with many appealing properties, although their widespread adoption has been hampered by the difficulty of applying inferential tools. In particular, the question "How reliable is the selection of those features?" has proved difficult to address, partially due to the complexity of defining a false discovery in the penalized regression setting. Here, I define a false inclusion as a variable that is independent of the outcome regardless of whether other variables are conditioned on. This definition permits straightforward estimation of the number of false inclusions when the correlation among predictors is mild. I will demonstrate the accuracy of the approach using simulated data and its practical utility using gene expression data from the Cancer Genome Atlas.

email: *patrick-breheny@uiowa.edu*

TOWARD PERSONALIZED PAN-OMIC ASSOCIATION ANALYSIS UNDER COMPLEX STRUCTURES

Eric P. Xing*, Carnegie Mellon University

A fundamental aim of modern medical genetics is to connect variations in clinical phenotypes with variations in the genome so that one can identify druggable genetic artifacts, predict clinical outcomes, and practice personalized medicine. The existing approaches for genetic analysis of complex human diseases remain inadequate in meeting many of the challenges toward this aim, such as, incorporating complex structural information to improve power; scaling up to ultra-high dimensionality to capture higher-order effects; adjusting the statistical model to allow personalizable inference; and furthermore, providing software and cloud API for easy computing. In this talk, I will discuss our recent efforts in developing mathematically rigorous, computationally tractable, and user-friendly tools for medical genetic inference and clinical prediction in presence of multiple confounders, rich prior knowledge, and needs for capturing both shared patterns and individual signatures in complex genetic effects. Our preliminary results promises to offer a practical basis for personalized medicine in the Big Data era of genomic healthcare.

email: *epxing@cs.cmu.edu*

88. CAUSAL INFERENCE

A CAUTIONARY TALE: MEDIATION ANALYSIS APPLIED TO CENSORED SURVIVAL DATA

Isabel R. Fulcher*, Harvard University

Eric J. Tchetgen Tchetgen, Harvard University

Paige L. Williams, Harvard University

Recent advances in causal mediation analysis have formalized conditions for estimating direct and indirect effects from empirical observations in the contexts of various models and outcomes. These approaches have been extended to a number of models for survival outcomes including the accelerated failure time (AFT) model. In this setting, it has been suggested that under standard assumptions, the "difference" and "product" methods produce equivalent estimates of the indirect effect of exposure on survival outcomes. We show that these two methods may produce substantially different estimates in the presence of censoring, due to a form of model misspecification. In simulation studies, we investigate implications of this phenomenon in estimating indirect effects for Normal and Weibull time-to-event outcomes. For normally-distributed survival outcomes with censoring, we show that the product and difference estimates are similar in large samples, but differ slightly in finite samples. Under the Weibull model, the difference method fails to be consistent for the indirect effect. We consider the implication of our findings in estimating the indirect effect of HIV status mediated through height for age at sexual maturity and the indirect effect of

combination treatment mediated through viral suppression on time to death or opportunistic infection among HIV-infected adults.

email: isabelfulcher@g.harvard.edu

SIMPLER APPROACH FOR MEDIATION ANALYSIS FOR DICHOTOMOUS MEDIATORS IN LOGISTIC REGRESSION

Hani Samawi, Georgia Southern University

Jingxian Cai*, Georgia Southern University

Harash Rochani, Georgia Southern University

Daniel Linder, Georgia Southern University

Mediation is a hypothesized causal chain in which one variable affects a second that, in turn, affects a third. It mediates the relationship between predictors and outcomes. To select and test for a potential mediator, the potential mediator should be associated with the predictor variable and with the outcome variable and should lie in the causal pathway between the predictor and the response The mediation analysis for continuous response variables is well developed in the literature and it can be shown that the total effect of(X)on(Y),c, is equal to c'+ab, where ab is the mediation effect of the variable(M).However, for categorical responses mediation analysis still not fully developed. In this paper, we propose and developed a new approach using the latent variable technique to adjust for c=c'+ab. Our intensive simulation study and theoretical developments showed that on average the proportion of the mediation effect of using our latent variable approach relative to direct approach is about 0.412. Real data example is used to illustrate the proposed approach.

email: xc00056@georgiasouthern.edu

PROPENSITY SCORE AND DOUBLY ROBUST METHODS FOR ESTIMATING THE EFFECT OF TREATMENT ON CENSORED COST

Jiaqi Li*, University of Pennsylvania

Nandita Mitra, University of Pennsylvania

The estimation of treatment effects on medical costs is complicated by the need to account for informative censoring, skewness and the effects of confounders. Since medical costs are often collected from observational claims data, we investigate propensity score (PS) methods such as covariate adjustment, stratification and inverse probability weighting taking into account informative censoring of the cost outcome. We compare these more commonly used methods to doubly robust estimation (DR). We then use a machine learning approach called Super-Learner (SL) to choose among conventional cost models to estimate regression parameters in the DR approach and to choose among various model specifications for PS estimation. Our simulation studies show that when the PS model is correctly specified, weighting and DR perform well. When the PS model is misspecified, the combined approach of DR with SL can still provide unbiased estimates. SL is especially useful when the underlying cost distribution comes from a mixture of different distributions or when the true PS model is unknown. We apply these approaches to a cost analysis of two bladder cancer treatments, cystectomy versus bladder preservation therapy, using SEER-Medicare data.

email: jiaqili@upenn.edu

SEMIPARAMETRIC EFFICIENT ESTIMATION OF COARSE STRUCTURAL NESTED MEAN MODELS IN THE PRESENCE OF INFORMATIVE CENSORING WITH APPLICATION TO THE EFFECT OF ONE-YEAR OF HAART

Shu Yang*, Harvard University

Judith Lok, Harvard University

Coarse Structural Nested Mean Models (SNMMs, Robins, 1998) provide useful tools to estimate treatment effects from longitudinal observational data. In this talk, we present the semiparametric efficient estimators for the estimation of the parameters of time-dependent coarse SNMMs, when the data are subject to informative censoring. Using semiparametric efficient theory, we show that the semiparametric efficient score belongs to a class of inverse probability of censoring weighted functions. The efficient score, and therefore the optimal weighted estimator, depends on unknown population quantities. We propose a locally semiparametric efficient estimator which simplifies calculation and which can achieve the semiparametric efficiency bound under a set of extra assumptions. We show that the proposed estimator is multiply robust under several combinations of nuisance models. The semiparametric approach is attractive because we do not need to specify the full likelihood, and multiple-robustness adds multiple protections on estimation against possible model misspecification. Simulation studies demonstrate that the proposed estimator performs well in finite samples. We illustrate the proposed estimator by investigating how the CD4 count increase due to one year of highly active antiretroviral treatment (HAART) depends on the time between HIV infection and HAART initiation in HIV-positive patients with early and acute infection.

email: yangshuyounggirl@gmail.com

A NEW WEIGHTED PARTIAL LIKELIHOOD METHOD FOR ESTIMATING MARGINAL STRUCTURAL HAZARD MODELS

Olli Saarela*, University of Toronto

Zhihui Liu, Cancer Care Ontario and University of Toronto

Parameters in marginal structural Cox models can be estimated through maximizing an inverse probability of treatment (IPT)

weighted Cox partial likelihood. Herein we propose an alternative weighted partial likelihood function for estimating parametric marginal structural hazard regression models, based on case-base sampling of person-moments, resulting in a weighted logistic regression form estimating function. The proposed method enables estimation of absolute hazards in addition to hazard ratios, and can accommodate continuous-time IPT weights. In terms of computational convenience, the proposed method resembles the discrete time pooled logistic regression method commonly used for estimating marginal structural Cox models, but works in continuous time, without rare disease approximations. We show that the resulting estimating equation is unbiased, study its properties through simulations, and illustrate the method in an application to modeling the effects of repeated treatment procedures in cancer patients. We also consider simulation of time-dependent confounding under a data-generating mechanism consisting entirely of continuous-time processes.

email: *olli.saarela@utoronto.ca*

IS THE DOUBLE ROBUST ESTIMATOR REALLY ROBUST?

Xavier de Luna*, Umeå University

Eva Cantoni, University of Geneva

In this talk we focus on situations where the interest lies in the estimation of a parameter, a functional from the distribution having generated the data (a random sample). We introduce semi-parametric estimators, which are able to deal simultaneously with two common challenges within this general context: (i) not all observations from the sample intended are available (incomplete data due to dropout, selection, potential outcomes), and (ii) some of the available observations may be contaminated (generated by a nuisance distribution, outliers). Under an assumption of ignorable missingness/selection, popular semi-parametric estimators of the parameter of interest are augmented probability weighted (AIPW, doubly robust) estimators. They use two auxiliary models, one for the missingness/selection mechanism, and another for an outcome of interest, both given observed covariates. AIPW estimators are then robust to misspecification of one of these two models (but not both simultaneously). We introduce versions of AIPW, which provide, moreover, robustness to contamination of the distribution of interest. Theoretical and finite sample results are provided. We motivate the need of robust AIPW estimators with a follow up study on BMI combining data from an intervention study and population wide record linkage data.

email: *xavier.deluna@stat.umu.se*

COVARIATE BALANCING IN PROPENSITY SCORE-BASED METHODS FOR OBSERVATIONAL STUDIES

Adin-Cristian Andrei*, Northwestern University

In biomedical research, the gold standard for comparing outcomes in two or more groups remains the randomized clinical trial (RCT). However, oftentimes it is not feasible to conduct RCTs for reasons ranging from medical ethics to financial considerations. An increasingly-popular alternative is to conduct observational studies instead. The propensity score (PS) plays a fundamental role in bias control in observational studies. When PS-matching is the analytical approach used, it is important to ensure that adequate baseline covariate balancing has been achieved. Based on both simulated and actual data, we discuss formal tools for covariate balancing assessment.

email: *aandrei@nm.org*

89. FUNCTIONAL DATA ANALYSIS

SINGLE-INDEX MODELS FOR FUNCTION-ON-FUNCTION REGRESSION

Guanqun Cao*, Auburn University

Lily Wang, Iowa State University

We propose a general framework for smooth regression of a functional response on multiple functional predictors, in which the mean of the response is related to the linear predictors via an unknown link function. This model provides as a good tool for dimension reduction in regression with multiple predictors and it is more flexible than functional linear models. Assuming that the functional predictors are observed at discrete points, we use B-spline basis functions to estimate the slope functions and the link function, and propose an iterative estimating procedure. Moreover, we devise uniform convergence rates of the proposed spline estimators, and construct asymptotic simultaneous confidence bands for the slope functions for inference. Our proposed method is illustrated by simulation studies.

email: *gzc0009@auburn.edu*

MULTIVARIATE MULTISCALE FUNCTIONAL DATA ANALYSIS

Andrew N. Potter*, University of Pittsburgh

Stewart J. Anderson, University of Pittsburgh

In many medical studies, the analysis of several patient or animal characteristic measured simultaneously over multiple time scales is a primary focus. For example, in cohorts of heart failure patients, several important cardiac characteristics are measured over a circadian cycle and moreover, changes in the circadian cycle itself

over time are noted. If the data are assumed to be generated by an underlying function,f(t), observed with noise, functional data analysis and time series techniques can be used. To analyze such longitudinal data, we introduce a novel functional multiscale model. Our proposed framework improves the ease of analysis by incorporating each time scale as an independent variable. Using data from a cohort of ventricular Assist Device patients, our proposed model incorporates both "fast time" (circadian) and "slow time" (longitudinal) components, represented by f(t) and f(s) respectively, in the characterization of the cardiac trajectories. We also develop a non-parametric bootstrap based inference technique for the population mean functions and other important features. Then, a new graphical representation of the data is introduced. Based on the results of simulations and analyses of real data, the new method appears to be more effective than existing techniques at detecting temporal changes in shape and features observed in patient cardiac outcomes.

email: *anp88@pitt.edu*

ORDINAL PROBIT WAVELET-BASED FUNCTIONAL MODELS FOR eQTL ANALYSIS

Mark J. Meyer*, Bucknell University

Jeffrey S Morris, University of Texas MD Anderson Cancer Center

Craig P. Hersh, Brigham and Women's Hospital

Jarrett D. Morrow, Brigham and Women's Hospital

Christoph Lange, Harvard School of Public Health

Brent A. Coull, Harvard School of Public Health

Current methods for conducting expression Quantitative Trait Loci (eQTL) analysis are limited in scope to a pairwise association testing between a single nucleotide polymorphism (SNPs) and expression probe set in a region around a gene of interest, thus ignoring the inherent between-SNP correlation. To determine association, p-values are then typically adjusted using Plug-in False Discovery Rate. As many SNPs are interrogated in the region and multiple probe-sets taken, the current approach requires the fitting of a large number of models. We propose to remedy this by introducing a flexible function-on-scalar regression that models the genome as a functional outcome. The model is formulated for a three-level ordinal categorical outcome in the Bayesian context and allows for the inclusion of a potentially large set of covariates. We examine the properties of the model in both simulation and in application to a chronic obstructive pulmonary disease genetic data set where eQTL analysis is of interest alongside a comparison of the standard approach.

email: *mark.john.meyer@gmail.com*

OPTIMAL DESIGN FOR SPARSE FUNCTIONAL DATA

So Young Park*, North Carolina State University

Luo Xiao, North Carolina State University

Jayson Wilbur, Metrum Research Group LLC

Ana-Maria Staicu, North Carolina State University

We consider an optimal design problem for sparse functional data. The primary objective is to find optimal sampling points for future data collection such that response can be most accurately predicted with the observations collected at those points. We formulate the problem as an optimization problem, and provide a unifying formulation for two major functional model frameworks: functional principal component analysis (FPCA) and functional linear model (FLM). We also propose a method for selecting number of optimal sampling points. Performance of the proposed method is thoroughly investigated via simulation study and application to real data example.

email: *spark13@ncsu.edu*

DETECTING OUTLIERS IN IMAGES OF DNA MOLECULES USING FUNCTIONAL DATA DEPTH AND MORPHOLOGICAL FEATURES

Subhrangshu Nandi*, University of Wisconsin, Madison

Alicia Nieto-Reyes, Universidad de Cantabria

Chengyue Wu, University of Science and Technology of China

Michael A. Newton, University of Wisconsin, Madison

When analyzing large-scale image data from single-DNA molecule measurements, it is important to identify outliers. There could be multiple reasons why an image of a molecule can be considered an outlier. Analysis of the morphological features of the molecules, via the analysis of the image pixel grey levels helps us identify some of them. A mathematical approach to detecting outliers in such images is by choosing the right functional data depth. We show that when these two approaches coincide they provide invaluable insights into the image processing of single molecule datasets.

email: *snandi@wisc.edu*

A BAYESIAN WAVELET BASED ANALYSIS OF LONGITUDINALLY OBSERVED SKEWED HETEROSCEDASTIC RESPONSE

Danisha S. Baker*, Florida State University

Eric Chicken, Florida State University

Debajyoti Sinha, Florida State University

Debdeep Pati, Florida State University

In this paper we propose a random effects based model for partial linear median regression function of a skewed longitudinal response using a wavelet expansion for the nonparametric part of the regression function. Parameters are estimated via a semiparametric Bayesian procedure using an appropriate Dirichlet process mixture prior for the skewed error distribution. Unlike common practices for

wavelet based regression for equally spaced data, we use a hierarchical mixture model as the prior for the wavelet coefficients. For the "vanishing" coefficients the model includes a level dependent prior probability mass at zero. This practice implements wavelet coefficient thresholding as a Bayes Rule. Consistency results have been obtained with only minor regularity conditions on the tail of the skewed and unimodal residual density. Practical advantages of our method are illustrated through a simulation study and via analysis of a cardiotoxicity study of children of HIV infected mother.

email: *dbaker@stat.fsu.edu*

90. HIGH DIMENSIONAL VARIABLE SELECTION

AN EFFICIENT METHOD FOR VARIABLE SELECTION IN LINEAR AND NONLINEAR MODELS

Arnab K. Maity*, Northern Illinois University

Sanjib Basu, Northern Illinois University

Appropriate model selection is a fundamental problem in the field of statistics. Models with large number of possible explanatory variables require special attention due to in-feasibility of huge model space. There are several suggestions available in the literature. Under the Bayesian approach, the classical way is to select the model with highest posterior probability. Using this fact the problem may be thought as a maximization problem over the model space where the objective function is the posterior probability of model and the maximization is taken place with respect to the models. We propose an efficient method for implementing this maximization and we illustrate its feasibility in high dimensional problem. By means of various simulation studies, this new approach has been shown to be efficient and to outperform other Bayesian methods namely median probability model and stochastic search variable selection. Theoretical justification has also been provided.

email: *arnabkrmaity@gmail.com*

COVARIANCE-INSURED SCREENING METHODS FOR ULTRAHIGH DIMENSIONAL VARIABLE SELECTION

Kevin He*, University of Michigan

Yi Li, University of Michigan

Ji Zhu, University of Michigan

Jiashun Jin, Carnegie Mellon University

Yanming Li, University of Michigan

Jian Kang, University of Michigan

Hyokyoung (Grace) Hong, Michigan State University

Effective screening methods are crucial to the analysis of big biomedical data. The popular sure independence screening relies on restricted assumptions such as the partial faithfulness condition, e.g, the partial correlation between outcome and covariates can be inferred from their marginal correlation. However, such a restrictive assumption is often violated, as the marginal effects of predictors may be quite different from their joint effects, especially when the covariates are correlated. We propose a covariance-insured screening (CIS) framework that utilizes the dependence among covariates and identify important features that are likely to be missed by marginal screening procedures such as sure independence screening. The proposed framework encompasses linear regression models, generalized linear regression models and survival models.

email: *kevinhe@umich.edu*

A DATA-DRIVEN APPROACH TO CONDITIONAL SCREENING OF HIGH DIMENSIONAL VARIABLES

Hyokyoung (Grace) Hong*, Michigan State University

Lan Wang, University of Minnesota

Xuming He, University of Michigan

Marginal screening is a widely applied technique to handily reduce the dimensionality of the data when the number of potential features overwhelms the sample size. Due to the nature of the marginal screening procedures, they are also known for their difficulty in identifying the so-called hidden variables that are jointly important but have weak marginal associations with the response variable. Failing to include a hidden variable in the screening stage has two undesirable consequences: (1) important features are missed out in model selection; and (2) biased inference is likely to occur in the subsequent analysis. Motivated by some recent work in conditional screening, we propose a data-driven conditional screening algorithm, which is computationally efficient, enjoys the sure screening property under weaker assumptions on the model, and works robustly in a variety of settings to reduce false negatives of hidden variables.

email: *hhong@stt.msu.edu*

SELECTION-ASSISTED SMOOTHED PARTIAL REGRESSION ESTIMATION AND INFERENCE FOR HIGH-DIMENSIONAL LINEAR MODEL

Zhe Fei*, University of Michigan

Yi Li, University of Michigan

Ji Zhu, University of Michigan

Variable selection and post-selection inference in high-dimensional data analysis has been widely studied in recent 20 years. In this project we propose a bagging estimator of coefficients in linear

model with more predictors than observations, the bagged estimator is derived from sub-sampling variable selection and partial regression of covariates so that each coefficient could be estimated whether it is selected or not. With mild sparsity and beta min condition, we prove that the proposed estimator is asymptotically unbiased and normally distributed. In addition, we provide the nonparametric delta-method estimate of standard deviation for the smoothed estimator according to Efron's 2014 paper. Thus we are able to derive confidence intervals and p-values for testing the significance of each and every regression coefficient. Our procedure provides estimation and inference for high-dimensional linear model similar to least square estimates (LSE) in low dimensional case. Simulations show the accuracy and coverage probablity of our method under different scenarios and compare with other methods including Van de Geer and Buhlmann's desparsified estimator of Lasso.

email: *feiz@umich.edu*

ON HIGH DIMENSIONAL INFERENCE

Qiang Sun*, Yale University

Heping Zhang, Yale University

We proposed a nuisance penalized regression framework for efficient inference for the parameter of interest, in the presence of high dimensional nuisance parameters. We provide scaling and complexity conditions under which an estimated nuisance parameter can be replaced by the true parameter without affecting the asymptotic efficiency of the parameter of interest. Theoretically, our framework provides the strongest inference guarantee under the weakest possible assumptions. Thorough numerical examples have been provided to back up our obtained methodology.

email: *qiang.sun@yale.edu*

A NEW CLASS OF MEASURES FOR TESTING INDEPENDENCE

Xiangrong Yin, University of Kentucky

Qingcong Yuan*, University of Kentucky

WITHDRAWN

We introduce a new class of measures for testing independence between two random vectors, which uses expected difference of conditional and marginal characteristic functions. In this paper, by choosing a particular weight function in the class, we propose a new index for measuring the dependence and study its property. To illustrate the use of such an index, one empirical version by slicing on one of the random vectors is developed. Its properties, asymptotics, connection with existing measures and applications in testing independence are discussed. Implementation and Monte Carlo results are also presented.

email: *qingcong.yuan@uky.edu*

DISTRIBUTED INFERENCE FOR HIGH DIMENSIONAL SEMI-PARAMETRIC ELLIPTICAL GRAPHICAL MODELS

Lu Tian*, University of Virginia

Pan Xu, University of Virginia

Quanquan Gu, University of Virginia

We propose a distributed communication-efficient inference method for semi-parametric Elliptical graphical models in the high dimensional regime. Our method distributes the data into k machines, and estimates the model parameter on each single machine using the data of size n/k. After the distributed estimation, our method averages the debiased estimators from k machines, and sparsifies the averaged estimator. We show that the resulting estimator attains the same rate as centralized estimation method, as long as k increases at a certain rate with respect to n. Thorough experiments on synthetic datasets backup our theory.

email: *lt2eu@virginia.edu*

91. NONPARAMETRIC METHODS

NOTES ON KERNEL BASED MODE ESTIMATION USING MORE EFFICIENT SAMPLING DESIGNS

Hani Samawi*, Georgia Southern University

Haresh Rochani, Georgia Southern University

JingJing Yin, Georgia Southern University

Daniel Linder, Georgia Southern University

Robert Vogel, Georgia Southern University

The mode estimation of a probability density function has become tractable in light of increasing computational power. The mode is one of the measures of the central tendency as well as the most probable value, which is not influenced by the tail of the distribution. Ranked set sampling (RSS) is a structural sampling method which improves the efficiency of parameter estimation in many circumstances and typically leads to a reduction in sample size and hence study cost. In this paper we investigate some of the asymptotic properties of kernel based mode estimation using RSS and compare it to mode estimation from that of simple random sampling (SRS). We demonstrate that kernel based mode estimation using RSS is consistent and asymptotically normal with lower variance than using SRS. Improved performance of the mode estimation using RSS compared to SRS is confirmed through a simulation study. A real data illustration using a Duchenne muscular dystrophy dataset is provided also.

email: *samawi.hani2@gmail.com*

AN EXACT TEST OF FIT FOR THE GAUSSIAN LINEAR MODEL USING OPTIMAL NONBIPARTITE MATCHING

Samuel D. Pimentel*, University of Pennsylvania

Dylan S. Small, University of Pennsylvania

Paul R. Rosenbaum, University of Pennsylvania

Beginning with Fisher, the fit of the Gaussian linear model has been examined with the aid of replicate or near-replicate observations. We refine this method in two ways. First, we construct near-replicates using an optimal nonbipartite matching that sets aside approximately 1/3 of the observations and pairs the remaining 2/3 of observations so that the total predictor distance within pairs is minimized. Second, we use the device employed in Tukey's one degree of freedom for nonadditivity to define a distance that focuses on predictors important to the model's predictions. Despite using the old fit to define the pairing, the test has exactly its stated level under the null hypothesis. A general problem with near-replicates is that they do not exist except when the dimensionality of the set of predicators is very low. The proposed method addresses dimensionality by betting that model failures will involve a subset of predictors that appear important in the fitted model. Simulations show the proposed method has reasonable power even when the set of predictors is padded with many predictors of no value. The test is demonstrated on a model for cost of care using patient data from the SUPPORT trial.

email: *spi@wharton.upenn.edu*

NON-INFERIORITY TEST BASED ON TRANSFORMATIONS FOR NON-NORMAL DISTRIBUTIONS

Santu Ghosh*, Georgia Regents 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 to a known reference drug by a small pre-specified amount. In this paper, we consider a three-arm non-inferiority trial consists of the placebo, a reference 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 test procedures for a three-arm non-inferiority trial based on transformations in conjunction with a normal approximation. Theoretical properties of our test methods are investigated. The effectiveness of our methods is illustrated through simulated data sets. Finally, a published clinical trial example is analyzed to demonstrate the benefits of the proposed test procedures.

email: *sghosh@gru.edu*

NONPARAMETRIC MULTIVARIATE CHANGE-POINT: ESTIMATION AND TESTING OF EXISTENCE

Sebastian J. Teran Hidalgo*, University of North Carolina, Chapel Hill

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

Michael C. Wu, Fred Hutchinson Cancer Research Center

In the set-up of the change-point problem, two sets of random vectors are obtained sequentially over time. At some unknown point in time, the relationship between these two vectors is believed to change, and it is of interest to estimate when this happens. The usual approach to this problem is to model the relationship between these two vectors, estimate two models for each time point, one for before and one for after the time point, and assess when the difference is largest between these pairs of models. In the current research, we propose a methodology to estimate the change-point without assuming a model. The method works for general data type and dimension. Also, because it is nonparametric, it can detect a wide range of changes in the relationship. This is accomplished by assessing nonparametrically the strength of the association between these two vectors before and after a change-point. Moreover, a test statistic is developed to test the hypotheses of nonexistence versus existence of a change-point. Theory shows consistency of the procedure. Simulations demonstrate correct Type-I error and Power. The method is computationally fast. We apply the method to a DNA methylation and copy number variants data sets.

email: *shidalgo@email.unc.edu*

ADJUSTED EMPIRICAL LIKELIHOOD METHOD FOR TREATMENT COMPARISONS IN LINEAR MODELS

Haiyan Su*, Montclair State University

Xi Kang, Montclair State University

Wei Ning, Bowling Green State University

In epidemiology and biomedical studies, comparisons of treatment effects in linear regressions are quite popular since the comparison controls other covariates through the regression model. Adjusted empirical likelihood method is an improvement of the empirical likelihood by adding a point to the profile likelihood. It has been shown to preserve all the asymptotic properties of the EL method. The coverage probability of the confidence interval fromAEL is also improved particularly when the sample size is small. Attracted by the nice properties of the AEL method, we propose an adjusted empirical likelihood-based method for comparing treatment effects in linear models in this study. We show that the test statistic follows chi-square distribution asymptotically. The numerical performance of the proposed method will be evaluated from numerical simulations.

email: *suh@mail.montclair.edu*

PENALISED SPLINE ESTIMATION FOR GENERALISED PARTIALLY LINEAR SINGLE-INDEX MODELS

Yuankun Zhang*, University of Cincinnati

Yan Yu, University of Cincinnati

Chaojiang Wu, Drexel University

Generalised linear models are frequently used in modeling the relationship of the response variable from the general exponential family with a set of predictor variables, where a linear combination of predictors is linked to the mean of the response variable. We propose a penalised spline (P-spline) estimation for generalised partially linear single-index models, which extend the generalised linear models to include non-linear effect for some predictors. The proposed models can allow flexible dependence on some predictors while overcome the "curse of dimensionality". We implement a P-spline profile likelihood estimation using the readily available R package mgcv, leading to straightforward computation. Simulation studies are considered under various link functions. In addition, we examine different choices of smoothing parameters. Simulation results and real data applications show effectiveness of the proposed approach. Finally, some large sample properties are established.

email: *zhangyk@mail.uc.edu*

92. SPATIOTEMPORAL MODELING

MODELLING NONLINEAR LAGGED EFFECTS WITH SPATIAL HETEROGENEITY

Lung-Chang Chien*, University of Texas School of Public Health, San Antonio

Kai Zhang, University of Texas School of Public Health, Houston

Yuming Guo, University of Queensland School of Public Health

Hwa-Lung Yu, National Taiwan University

In environmental health research, the influence of exposures not only affect current health, but these exposures can also have a lagged effect, distributing over several subsequent response measures for a certain period of time. In addition, locational effects, which can manifest as spatial heterogeneity, may also exist in both exposures and health outcomes. For the purpose of considering a nonlinearity association between lagged effects and health outcomes, recent studies more likely applied the distributed lag nonlinear model (DLNM) to elaborate complex relationships among health outcomes, exposures and lag via a cross-basis function. However, there is a gap in the DLNM to analyze geographic information data. An extensive DLNM containing a spatial function from the Markov random fields is proposed and applied in two environmental health studies for displaying how this modeling approach can analyze geographical data and nonlinear lagged effects simultaneously. The two applied studies show significantly high-risk areas where people living there may be more likely vulnerable to diseases after controlling for both linear demographics and socioeconomic effects and nonlinear lagged effects. The visualization of the spatial function in the DLNM can also carry out with mapping techniques.

email: *Lung-Chang.Chien@uth.tmc.edu*

STEPWISE AND STAGEWISE APPROACHES FOR SPATIAL CLUSTER DETECTION

Jiale Xu, University of Wisconsin

Ronald Gangnon*, University of Wisconsin

Spatial cluster detection is an important tool in many areas such as sociology, botany and public health. Previous work has mostly taken either hypothesis testing framework or Bayesian framework. In this paper, we propose a few approaches under a frequentist variable selection framework for spatial cluster detection. The forward stepwise methods search for multiple clusters by iteratively adding currently most likely cluster while adjusting for the eects of previously identified clusters. The stagewise methods also consist of a series of steps, but with tiny step size in each iteration. We study the features and performances of our proposed methods using simulations on idealized grids or real geographic area. From the simulations, we compare the performance of the proposed methods in terms of estimation accuracy and power of detections. These methods are applied to the the well-known New York leukemia data as well as a Midwest States poverty data.

email: *ronald@biostat.wisc.edu*

A SPATIO-TEMPORAL APPROACH FOR MODELING THE EFFECTS OF WEATHER AND CLIMATE ON MALARIA DISTRIBUTIONS IN WEST AFRICA

Ali Arab*, Georgetown University

Monica Jackson, American University

Cezar Kongoli, University of Maryland, and National Oceanic and Atmospheric Administration (NOAA), National Environmental Satellite Data and Information Service (NESDIS)

WITHDRAWN

Malaria is a leading cause of mortality worldwide. There is currently conflicting data and interpretation on how variability in climate factors affects the incidence of malaria. In this work, we present a hierarchical Bayesian modeling framework for the analysis of malariaversusclimate factors in West Africa. The proposed hierarchical Bayesian framework takes into account spatio-temporal dependencies, and is applied to annual malaria and climate data from ten West African countries (Benin, Burkina Faso, Côte d'Ivoire, Gambia, Ghana, Liberia, Mali, Senegal, Sierra Leone, and Togo) during the period 1996-2006. Our results show a statistically significant

correspondence between malaria rates and the climate variables considered. The two most important climate factors are found to be average annual temperature and total annual precipitation, and they show negative association with malaria incidence. This modeling framework provides a useful approach for studying the impact of climate variability on the spread of malaria and may help to resolve some conflicting interpretations in the literature.

email: *ali.arab@georgetown.edu*

AN UNCERTAINTY QUANTIFICATION APPROACH FOR DETERMINISTIC SPATIAL INTERPOLATIONS

Robert J. Waken*, Baylor University

Soohyun Kwon, Kyungpook National University

GyuWon Lee, Kyungpook National University

Joon Song, Baylor University

Deterministic spatial interpolators, like inverse distance weighting (IDW) and regression-based inverse distance weighting (RIDW), offer flexible and fast mean modeling alternatives in spatial analysis to their stochastic counterparts, but exhibit a distinct disadvantage due to the inability to properly assess uncertainty.We propose a flexible stochastic uncertainty attachment scheme for spatial/spatiotemporal deterministic interpolators through measurement error modeling. The proposed method is applied to radar rainfall estimation and compared with some existing methods.

email: *rj_waken@baylor.edu*

MODELING HIGH DIMENSIONAL MULTICHANNEL ELECTROENCEPHALOGRAMS

Lechuan Hu*, University of California, Irvine

Hernando Ombao, University of California, Irvine

The goal of this paper is to develop a procedure for fitting a vector autoregressive (VAR) model to a high dimensional multi-channel electroencephalogram (EEG) data. From the parameter estimates of the VAR model, we obtain connectivity measures with the intended clinical application of identifying connectivity measures (for specific pairs or between groups of channels) as predictors for a stroke patient's ability to regain motor functionality. The key step here is to fit a high dimensional VAR model which is a non-trivial task. A VAR of order d with P channels require d*p^2 number of parameters. We develop a new two-step computational procedure to analyze high-dimensional EEG signals under VAR framework. Our method is distinguished from both specificity of non-connectivity and sensitivity of connectivity strength compared with the LSE and LASSO methods. Moreover, we present visualization results to help quickly identify the functional connectivity between channels in a resting-state EEG data.

email: *lechuanh@uci.edu*

NON-SEPARABLE DYNAMIC NEAREST-NEIGHBOR GAUSSIAN PROCESS MODELS FOR LARGE SPATIO-TEMPORAL DATA WITH AN APPLICATION TO PARTICULATE MATTER ANALYSIS

Abhirup Datta*, University of Minnesota

Sudipto Banerjee, University of California, Los Angeles

Andrew O. Finley, Michigan State University

Nicholas A.S. Hamm, University of Twente

Martijn Schaap, TNO Built Environment and Geosciences

Particulate matter (PM) is a class of malicious environmental pollutants known to cause detrimental effects on human health. Regulatory efforts aimed at curbing PM levels in different countries require high resolution space-time maps that can identify red-flag regions exceeding statutory concentration limits. Continuous space-time Gaussian Process (GP) models can potentially deliver uncertainty quantified map predictions for PM levels. However, traditional GP based approaches are thwarted by computational challenges posed by large datasets. We construct a novel class of scalable Dynamic Nearest Neighbor Gaussian Process (DNNGP) models that can provide a sparse approximation to any non-separable and possibly non-stationary spatio-temporal GP. The DNNGP can be used as a sparsity-inducing prior for spatio-temporal random effects in any Bayesian hierarchical model to deliver full posterior inference. Storage and memory requirements for a DNNGP model are linear in the size of the dataset thereby delivering massive scalability without sacrificing inferential richness. Extensive numerical studies reveal that the DNNGP provides substantially superior approximations to the underlying process than low rank approximations. Finally, we use the DNNGP to analyze a massive air quality dataset to considerably improve predictions of PM levels across Europe in conjunction with the LOTOS-EUROS chemistry transport models (CTMs).

email: *datta013@umn.edu*

AN EXPLORATORY COHERENCE ANALYSIS OF ELECTROENCEPHALOGRAMS USING THE FUNCTIONAL BOXPLOTS APPROACH

Duy Ngo*, University of California, Irvine

Hernando Ombao, University of California, Irvine

Many model-based methods have been developed over the last several decades for analysis of electroencephalograms (EEG) in order to understand electrical neural data. In this work, we propose to study the spectral properties of brain signals and the cross-oscillatory interactions between brain activity at different channels using functional boxplots. The functional boxplot approach produces a median curve of coherence estimate which tells us, on the average, the behavior of dependence between pairs of channels. The functional median curve is not obtained by merely connecting the

medians of boxplots and thus gives an estimated curve that is truly representative of the curves across many trials. Many model-based methods produce estimate coherence curves that are based on the average and hence could be adversely affected by outliers. On the contrary, our approach is robust to unusual coherence curves. In addition, this approach identifies a functional median, summarizes variability and detects potential outliers. By using rank-based non-parametric tests, we also investigate the stationarity of EEG traces across an exam acquired during resting-state by comparing the Fisher z-transformation coherence during the early vs. late phases of a single resting-state EEG exam.

email: *dngo5@uci.edu*

93. SURVIVAL ANALYSIS: MULTIVARIATE AND HIERARCHICAL

CIRCULATORY DISEASE MORTALITY IN A POOLED ANALYSIS OF THE MASSACHUSETTS AND CANADIAN TUBERCULOSIS FLUOROSCOPY COHORTS

Van Tran*, National Cancer Institute, National Institutes of Health

Lydia B. Zablotska, University of California, San Francisco

Alina V. Brenner, National Cancer Institute, National Institutes of Health

Mark P. Little, National Cancer Institute, National Institutes of Health

Although there is strong evidence that exposure to high-dose ionizing radiation is associated with elevated risk for circulatory diseases, there is no consensus on the effect of fractionated low to moderate doses. Previous studies have analyzed circulatory disease in two datasets of persons given repeated fluoroscopic exposures as part of the diagnosis and treatment for tuberculosis in Massachusetts (13,568 patients between 1916 and 1961) and in Canada (63,707 patients between 1950 and 1987). Pooling data from both cohorts, we fitted Poisson regression models to estimate excess relative risk (ERR) per Gy from cumulated x-ray fluoroscopy dose. There was no excess mortality risk for all circulatory diseases (n=12,545; ERR/Gy=-0.0024; 95% CI -0.0072, 0.0027;p=0.352). Adjusting for years since last exposure and age at first exposure (p= 0.007) slightly increased excess risk (ERR/Gy= -0.0014; 95% CI -0.0053, 0.0014). For non-ischemic heart disease, ERR was slightly decreased at higher doses (n=1,674; ERR/Gy=-0.0111; 95% CI -0.0221, 0.0016;p=0.085). Dose fractionation, years since last exposure and age at first exposure were generally found not to modify excess risk. Under a unified analysis of both cohorts, there was no indication of excess mortality from circulatory diseases associated with fractionated low-to-medium doses of ionizing radiation.

email: *thanh.tran@nih.gov*

A JOINT FRAILTY MODEL FOR ZERO-INFLATED RECURRENT EVENTS AND A TERMINAL EVENT IN A MATCHED STUDY

Cong Xu*, The Pennsylvania State University

Ming Wang, The Pennsylvania State University

Vernon Chinchilli, The Pennsylvania State University

Recurrent events or repeated events occur frequently during the follow-up in longitudinal clinical studies. In practice, if the recurrence is some severe event, the censoring time caused by death is highly likely to be informative. Our work was motivated by the Assessment, Serial Evaluation, and Subsequent Sequelae of Acute Kidney Injury (ASSESS-AKI) Consortium. In this multi-center matched cohort study, the individuals with AKI and non-AKI during hospitalization are matched on major baseline confounders based on a prioritized set of criteria. A certain large proportion of subjects who have no recurrent AKI being observed manifest the "zero-inflated" nature of the data. Also, there is no probability of being "cured" when the terminal event is death. Thus, a joint frailty model for zero-inflated recurrent events and death is proposed with a matched logistic model for zero recurrent event in this matched cohort. We incorporate two frailties to measure the dependency between AKI and non-AKI subjects within a matched pair and that among recurrent AKI events within one individual. By sharing the frailties, death may be dependent on recurrent AKI event history. Maximum likelihood estimation and inference are obtained based on a Monte Carlo EM algorithm. Extensive simulation studies are provided.

email: *congxu@hmc.psu.edu*

A THREE-STATE MARKOV FRAILTY MODEL FOR INTERVAL CENSORED CARIES LIFE HISTORY DATA

Daewoo Pak*, Michigan State University

Chenxi Li, Michigan State University

David Todem, Michigan State University

In this paper, we propose a three-state frailty Markov model coupled with a likelihood-based inference to analyze tooth-level life course data in caries research. This analysis proves challenging because of intra-oral clustering, interval censoring, multiplicity of caries states, and computational complexities. We develop a Bayesian approach to predict future caries transition probabilities given observed life-history data. Numerical experiments demonstrate that the proposed methods perform very well in finite samples with moderate sizes. The practical utility of the model is illustrated using life course data from a unique longitudinal study of dental caries in young low-income urban African-American children. In this analysis, we evaluate for any spatial symmetry in the mouth with respect to the life course of dental caries, and whether the same type of tooth has a similar decay process in boys and girls.

email: *pakdaewo@msu.edu*

A SCORE TEST FOR COPULA-BASED BIVARIATE SURVIVAL MODEL, WITH AN APPLICATION TO GENOME-WIDE ANALYSIS FOR PROGRESSION OF AGE-RELATED MACULAR DEGENERATION

Yi Liu*, University of Pittsburgh

Ying Ding, University of Pittsburgh

Wei Chen, University of Pittsburgh

Motivated by a genome-wide study in identifying genetic causes for progression of Age-Related Macular Degeneration (AMD), we propose a score test for bivariate time-to-event data through a copula model. Specifically, the progression times for both eyes are jointly modeled through a copula function with a common (individual-level) genetic effect. We derive the theoretical form of the score test and develop an efficient computational approach under the proportional hazards marginal distributions. Extensive simulation studies are conducted under the genome-wide setting with various marginal distributions, censoring schemes and association strength to evaluate the test performance. Our proposed score test shows great advantages in convergence stability and computational efficiency as compared to the likelihood ratio test and the Wald test while producing very similar power and type I error. When compared with the marginal model under the "working independence", the score test yields an improved type I error control. We apply our method on two large randomized clinical trials, Age-Related Eye Disease Studies (AREDS) and AREDS2, in identifying novel genetic variants for the progression of AMD. We show that the proposed method is computationally feasible on a whole-genome scale and leads to potentially valuable findings for clinical practice.

e-mail: *yil127@pitt.edu*

MIXTURE MODELS FOR LEFT-CENSORED AND IRREGULARLY-CENSORED DATA: APPLICATIONS TO A CANCER SCREENING COHORT ASSEMBLED FROM ELECTRONIC HEALTH RECORDS

Li C. Cheung*, George Washington University

Qing Pan, George Washington University

Noorie Hyun, National Cancer Institute, National Institutes of Health

Barbara Fetterman, Kaiser Permanente, Northern California

Philip E. Castle, Albert Einstein School of Medicine

Hormuzd A. Katki, National Cancer Institute, National Institutes of Health

For cost-effectiveness and efficiency, many large-scale general-purpose cohort studies are being assembled within large health-care providers who use electronic health records. Two key features of such data are that incident disease is interval censored between irregular visits and preexisting (prevalent) disease is left-censored. Because prevalent disease is not always immediately diagnosed, some disease diagnosed at future visits is actually undiagnosed prevalent disease. We demonstrate that the naive Kaplan-Meier cumulative risk estimator underestimates risks at early times and overestimates later risks. We propose a general family of mixture models for interval-censored incident disease and left-censored prevalent disease that we call prevalence-incidence models. Parameters for parametric prevalence-incidence models, such as the logistic regression and Weibull survival model (logistic-Weibull) are estimated by an EM algorithm. When there are no covariates, we show how to calculate cumulative risk non-parametrically. We compare naive Kaplan-Meier, logistic-Weibull, and non-parametric estimates of cumulative risk in the cervical cancer screening program at Kaiser Permanente Northern California. Our findings support use of logistic-Weibull models to develop the risk estimates that underlie current U.S. cervical cancer screening guidelines.

e-mail: *licheung@gwmail.gwu.edu*

REGRESSION ANALYSIS OF INTERVAL CENSORED DATA IN THE PRESENCE OF CURED SUBGROUP AND MISMEASURED COVARIATES

Yeqian Liu*, University of Missouri, Columbia

Tao Hu, Capital Normal University

Jianguo Sun, University of Missouri, Columbia

We discuss regression analysis of interval-censored data, a commonly occuring type of failure time data, arising from the proportional hazards model. Furthermore, there may exist a cured subgroup, meaning that a proportion of study subjects are not susceptible to the failure event of interest. We also consider the case where one or more explanatory variables in the model are subject to measurement error. This error should be taken into account in the estimation of the model, to avoid biased estimators. A general approach that exists in the literature is the SIMEX algorithm, a method based on simulations which allows one to estimate the effect of measurement error on the bias of the estimators and to reduce this bias. We extend the SIMEX approach to the mixture cure model with interval censored data.

e-mail: *yldg5@mail.missouri.edu*

ACCOUNTING FOR HETEROGENEITY WHEN EVALUATING SURROGATE ENDPOINTS IN A DISCRETE-TIME SURVIVAL MODEL

Andrew J. Spieker*, University of Washington

Ying Huang, Fred Hutchinson Cancer Research Center

There is a great interest in developing a vaccine for protection against HIV. Proper measures of clinical efficacy such as time to infection can be time consuming and costly to obtain; hence, identification of surrogate endpoints such as vaccine induced immune

response biomarkers would be of great value to screen out ineffective vaccines without demanding large samples. Repeated low-dose challenge experiments have been proposed as a realistic model for primate studies in order to determine surrogate value of immune biomarkers in HIV vaccine trials. A discrete failure time model has been proposed for evaluating potential surrogate endpoints via measures such as the proportion associative effect statistic, which has good performance when model assumptions hold. However, heterogeneity across study subjects can meaningfully impact inference for such measures, producing potentially underpowered study design and misleading preliminary results about vaccine efficacy and the biomarker's surrogate value. Given that homogeneity across subjects is almost assuredly not satisfied in practice, there is an unmet need to address this assumption in order to design studies with adequate power and provide valid inference in data analysis. In this talk, we will present a proposed model extension to account for heterogeneity using subject-specific random effects.

e-mail: *ajspiek@uw.edu*

94. SOME NEW DEVELOPMENTS IN THE MODERN LONGITUDINAL DATA ANALYSIS

MARGINAL REGRESSION MODEL FOR LONGITUDINAL NETWORK DATA

Yan Zhou, Merck & Co.

Peter X.K. Song*, University of Michigan

Longitudinal network data refer to temporal measurements repeatedly collected from units/subjects on a network. Such data arise frequently from studies in social and health sciences. We develop a new regression methodology to account for both within-subject and network-level correlations in the context of marginal models for continuous and discrete outcomes. To utilize both prior network topology and data-driven network correlation into the regression analysis, we propose a hybrid estimating function approach for statistical estimation and inference. Moreover, a Godambe information based tuning strategy is proposed to allocate hybrid weights so that the resulting estimation achieves optimal efficiency. A clear advantage of the proposed estimation method is its computation feasibility. Also, it has desirable large-sample properties in both estimation and inference. The proposed estimation method is evaluated through simulation studies and illustrated by a real example of neuroimaging data concerning an association study of iron deficiency on infant's auditory recognition memory.

email: *pxsong@umich.edu*

MULTIVARIATE SEMI-CONTINUOUS TWO PART FIXED EFFECTS MODELS

Yaoguo Xie, University of Wisconsin, Madison

Zhengjun Zhang*, University of Wisconsin, Madison

Paul Rathouz, University of Wisconsin, Madison

Bruce Barrett, University of Wisconsin, Madison

Semi-continuous data, also known as zero-inflated continuous data, have a substantial portion of responses equal to a single value (typically 0) and a continuous, right-skewed distribution among the remaining positive values. For joint modeling of two clustered semi-continuous responses, a bivariate two part fixed effects model is proposed. The covariate effects in the two positive parts of the model can be constrained to be proportional to the covariate effects in the logistic part. When this assumption obtains, it is shown that, both theoretically and experimentally, the constrained model will be more efficient than the unconstrained model and thus provide a deeper understanding of the data structure. The proposed model is applied to data from a randomized controlled trial, which evaluated potential preventive effects of meditation or exercise on duration and severity of acute respiratory infection (ARI) illness.

email: *zjz@stat.wisc.edu*

GENERALIZED ADDITIVE PARTIAL LINEAR MODELS FOR CLUSTERED DATA WITH DIVERGING NUMBER OF COVARIATES USING GEE

Hua Liang*, George Washington University

Heng Lian, University of New South Wales, Australia

Lan Wang, University of Minnesota

We study flexible modeling of clustered data using marginal generalized additive partial linear models with diverging number of covariates. Generalized estimating equation based estimators are derived after we approximate the nonparametric functions by polynomial splines. We establish the asymptotic properties in a "large n, diverging p" framework. More specifically, we establish the consistency and asymptotic normality of the estimators for the linear parameters under mild conditions. We further propose penalized estimating equations based procedure for variable selection, which can identify non-zero components to obtain the final selection and estimation results. The proposed variable selection procedure is shown to have the oracle property and allows the number of parameters in the linear part to increase at the same order of the sample size under some general conditions. Extensive Monte Carlo simulations demonstrate that the proposed methods work well with moderate sample sizes. A real dataset is analyzed as an illustration.

email: *hliang@gwu.edu*

THE MODELING OF MEDICAL EXPENDITURE DATA FROM A LONGITUDINAL SURVEY USING THE GENERALIZED METHOD OF MOMENTS (GMM) APPROACH

Zachary Hass, Purdue University

Michael Levine*, Purdue University

Laura P. Sands, Virginia Tech

Jeffrey C.-Y. Ting, American Credit Acceptance

Huiping Xu, Indiana University Purdue University, Indianapolis

Medical expenditure data analysis has recently become an important problem in biostatistics. These data typically have a number of features making their analysis rather difficult. Commonly, they are heavily right-skewed, contain a large percentage of zeros and often exhibit large numbers of missing observations due to death and/or the lack of follow-up. They are also commonly obtained from records that are linked to large longitudinal data surveys. In this manuscript, we suggest a novel approach to modeling these data through the use of GMM (Generalized Method of Moments) estimation procedure combined with appropriate weights that account for both dropout due to death and the probability of being sampled from among National Long Term Care Survey (NLTCS) subjects. This approach seems particularly appropriate due to the large number of subjects relative to the length of observation period (in months). We also use a simulation study to compare our proposed approach with and without the use of weights. The proposed model is applied to medical expenditure data obtained from the 2004-2005 NLTCS linked Medicare data base. The results suggest that the amount of medical expenditures incurred is strongly associated with higher number of activities of daily living (ADL) disabilities and self-reports of unmet need for help with ADL disabilities.

email: *mlevins@purdue.edu*

95. STATISTICAL CONSIDERATIONS IN PERSONALIZED MEDICINE: CONCEPT AND METHODOLOGY

THE IMPACT OF COMPANION DIAGNOSTIC DEVICE MEASUREMENT PERFORMANCE ON CLINICAL VALIDATION OF PERSONALIZED MEDICINE

Meijuan Li*, U.S. Food and Drug Administration

Tinghui Yu, U.S. Food and Drug Administration

Yun-Fu Hu, U.S. Food and Drug Administration

A key component of personalized medicine is companion diagnostics that measure biomarkers e.g. protein expression, gene amplification or specific mutations. Most of the recent attention concerning molecular cancer diagnostics has been focused on the biomarkers of response to therapy, such as KRAS mutations in metastatic colorectal cancer, EGFR mutations in advanced Non-small cell lung cancer, and BRAF mutations in metastatic malignant melanoma. The presence or absence of these markers is directly linked to the response rates of particular targeted therapies with small-molecule kinase inhibitors or antibodies. Therefore, testing for these markers has become a critical step in the target therapy of the above-mentioned tumors. The core capability of personalized medicine is the CDx's ability to accurately and precisely stratify patients by their likelihood of benefit (or harm) from a particular therapy. There is no reference in the literature discussing the impact of device's measurement performance e.g.analytical accuracy and precision on treatment effects, variances, and sample sizes of clinical trial for the personalized medicine. In this paper, using both analytical and estimation method, we assessed the impact of CDx measurement performance as a function of positive and negative predictive values and imprecision (standard deviation) on treatment effects, variances of clinical outcome, and sample sizes for the clinical trials.

e-mail: *meijuan.li@fda.hhs.gov*

PERSONALIZED ONCOLOGY IN 2015: NEW PARADIGMS IN CLINICAL TRIAL METHODOLOGY

Richard Macey Simon*, National Cancer Institute, National Institutes of Health

Personalized medicine has progressed more rapidly in the treatment of cancer patients than in other areas of medicine. Cancers are diseases of DNA dis-function and new diagnostic classification systems based on somatic genomic alterations are rapidly replacing traditional systems based on primary site and histology. A large proportion of the cancer drugs that have been approved by regulatory authorities in the past decade have an intended use for a restricted subset of patients. The intended use subset is often characterized by de-regulation of a gene related to the molecular target of the drug. Much current drug development in oncology involves co-development of a companion in-vitro diagnostic test for selecting the subset of patients who are likely to benefit from the drug. The companion diagnostics are often based on DNA sequencing of patients' tumors. Progress in the development of effective drugs has increased in oncology. The progress has been based on use of non-traditional clinical trial designs such as enrichment designs in which a relatively narrow subset of patients are selected for randomization instead of the usual broad eligibility trials. Adaptive enrichment designs and run-in designs have been developed for settings where a single candidate predictive biomarker is not known a-priori. There has also been increased use of prospective-retrospective designs in which a focused analysis of a single candidate predictive biomarker is performed using a previ-

ously conducted randomized clinical trial. There is also increased interest in carefully structured historically controlled clinical trials for cases where the new drug appears so effective in phase II trials that equipoise for conducting a phase III trials lost. The designs that have been used effectively in developing personalized oncology will be reviewed and some of the new design paradigms developed for current and future studies will be discussed.

e-mail: *rsimon@mail.nih.gov*

BIAS CORRECTION IN ESTIMATING THE HETEROGENEOUS TREATMENT EFFECT IN SUBGROUP ANALYSIS

Lu Tian*, Stanford University

Lee-Jen Wei, Harvard School of Public Health

One criticism of the simple subgroup analysis is the potential bias caused by imbalance in important baseline characteristics due to imperfect randomization. The bias can be severe when we attempt to estimate the covariate and treatment interaction, since the sample size of the subgroup is often small. A common approach to correct this bias is via fitting a multivariate regression model. While it is easy to implement, the validity of this model-based approach depends on the validity of the model assumptions. More importantly, when a nonlinear model is used to make the adjustment such as the Cox model in survival analysis, the estimator for the interaction in general does not converge to the target interaction when the sample size goes to infinity and the randomization becomes perfect. In this talk, we proposed a model-free approach for bias correction in studying treatment and covariate interactions to overcome the aforementioned inconsistency. We also propose to use restricted mean survival time rather than the hazard function to measure the interaction when the time to event outcome is of interest. The proposal is evaluated via numerical study as well as real data example.

e-mail: *lutian@stanford.edu*

THE BRAVE NEW WORLD OF CANCER CLINICAL TRIALS: LEARNING WHO BENEFITS FROM WHAT?

Donald Berry*, University of Texas MD Anderson Cancer Center

Cancer biologists are changing our understanding of the disease in amazing ways. Cancer clinical trialists are struggling to accommodate. Biomarkers enter the picture in two distinct ways. Baseline biomarkers are used to partition diseases into treatment "signatures." Longitudinal biomarkers measure the extent of disease post- treatment. Both types require statistical modeling, which I will describe. And both are fundamentally important for incorporating into clinical trial designs, a process that I will also describe.

e-mail: *dberry@mdanderson.org*

96. INNOVATIVE TECHNIQUES TOWARDS SOLVING THE COMPLEXITIES OF BIOMARKER DISCOVERY

DESIGNING DISEASE ELIMINATION STRATEGIES USING MODELS AND DATA FROM MULTIPLE SOURCES

John M. Marshall*, University of California, Berkeley

With an ever-growing quantity of data, decisions related to disease control are increasingly evidence-based, and models are frequently being used to synthesize data from multiple sources. As diseases approach low prevalence, transmission becomes increasingly focal and quantifying the heterogeneities in disease transmission becomes important for designing cost-effective elimination strategies. Challenges include quantifying heterogeneity, efficiently estimating parameters and weighting evidence. We discuss these challenges in the context of designing cost-effective malaria elimination strategies in low prevalence settings. In these settings, novel intervention delivery strategies are currently being designed including: a) reactive case detection, in which cases are detected passively; but upon detection, household members and neighbors are screened and treated as well; and b) network targeting, in which infected individuals who have recently traveled help to recruit individuals from their social network who have a relatively high likelihood of also carrying an imported infection. Exploring the benefits of these delivery strategies using models requires a quantitative understanding of disease transmission, the natural history of the malaria parasite, the ecology of the mosquito vector, environmental and seasonal determinants of risk, and the parameters and constraints of intervention delivery. Within this context, we describe an adventure in data integration.

e-mail: *john.marshall@berkeley.edu*

ANALYSIS OF PROPORTIONAL HAZARDS MODEL WITH SPARSE TIME-DEPENDENT COVARIATES

Jason Fine*, University of North Carolina, Chapel Hill

Regression analysis of censored failure observations via the proportional hazards model permits time-varying covariates which are observed at death times. In practice, such longitudinal covariates are typically sparse and only measured at infrequent and irregularly spaced follow-up times. Full likelihood analyses of joint models for longitudinal and survival data impose stringent modelling assumptions which are difficult to verify in practice and which are complicated both inferentially and computationally. We propose a simple kernel weighted score function which is valid under minimal assumptions. Two scenarios are considered: half kernel estimation in which observation ceases at the time of the event and full kernel estimation for data where observation may continue after the event, as with recurrent events data. It is established that these estimators are consistent

and asymptotically normal. However, they converge at rates which are slower than the parametric rates which may be achieved with fully observed covariates, with the full kernel method achieving an optimal convergence rate which is superior to that of the half kernel method. Simulation results demonstrate that the large sample approximations are adequate for practical use and may yield improved performance relative tolast value carried forward approach and joint modelling method. The analysis of data from a cardiac arrest study demonstrates the utility of the proposed methods.

e-mail: *jfine@email.unc.edu*

A MULTI-STEP CLASSIFIER IDENTIFIES COHORT HETEROGENEITY IN CANCERS LEADING TO IMPROVED ACCURACY OF PROGNOSTIC BIOMARKERS

Samuel Mueller*, University of Sydney

Ellis Patrick, Harvard Medical School

Jean Yang, University of Sydney

Ongoing research in cancers continues to highlight the extensive genetic diversity both within and between tumors. This intrinsic heterogeneity poses one of the central challenges in predicting patient clinical outcome as well as the personalisation of treatments. Efforts to classify patients according to disease trajectory have been vast and varied, and, despite advances in classification methods, moderate error rates in the accuracy of patient classification persist. That is to say there are subsets of patients who remain outside the sensitivity of the models proposed to date. Recent work, examining the prognostic ability of biomarkers to classify individual patients, provides a way of identifying the presence of distinct prognostic signals in a cohort. We demonstrate in three independent cancer cohorts (melanoma, breast, and colorectal) that clinico-pathologic variables can predict patients that are correctly classified by a biomarker; using this information in a multi-step classification procedure not only improves classification performance but also points to specific clinical attributes which can explain the heterogeneity in each cohort. This paves the way for a new generation of interpretable prognostic biomarkers for complex disease.

e-mail: *samuel.mueller@sydney.edu.au*

COX REGRESSION WITH EXCLUSION FREQUENCY-BASED WEIGHTS TO IDENTIFY NEUROIMAGING MARKERS RELEVANT TO HUNTINGTON'S DISEASE ONSET

Tanya P. Garcia*, Texas A&M University

Samuel Mueller, University of Sydney

Biomedical studies of neuroimaging and genomics have evolved into collecting large amounts of data on a small subset of subjects so as to not miss any informative features. An important goal in such studies is parsing through the data to identify truly informative features which provide better visualization of the data and are generally more cost-effective measures in future clinical trials. The goal becomes challenging, however, when the features are naturally interrelated and the response is a failure time prone to censoring. We propose to handle these inherent challenges through a novel variable selection technique which casts the problem into several smaller dimensional settings and extracts from this intermediary step the relative importance of each feature through data-driven weights called exclusion frequencies. The exclusion frequencies are used as weights in a weighted Lasso and shown to yield low false discovery rates and high geometric mean of sensitivity and specificity. We illustrate the method's advantages over existing ones in an extensive simulation study and apply the method to a neuroimaging study of Huntington's disease to identify those brain features most relevant to age of disease onset.

e-mail: *tpgarcia@sph.tamhsc.edu*

97. NEW DEVELOPMENTS OF BAYESIAN METHODS FOR CAUSAL INFERENCE

A NONPARAMETRIC BAYESIAN APPROACH FOR ESTIMATING THE AVERAGE CAUSAL EFFECT

Bo Lu, The Ohio State University

Steven N. MacEachern*, The Ohio State University

Ling Wang, The Ohio State University

Inferring a causal relationship is an important task in health and medical research. In an observational study, unlike a randomized experiment, treatment assignment is likely to be confounded with many factors. Under the potential outcome framework, propensity score based matching, stratification, and weighting approaches are commonly used to estimate the average treatment effect. We propose the use of Bayesian modeling in conjunction with the propensity score to estimate the causal effect. Flexible Bayesian models based on the Gaussian process are used to estimate the response surface for treatment and control groups. Conventional matching estimators can be reproduced through consideration of a limit of prior distributions. Moreover, nonparametric Bayesian techniques can be easily adapted to estimate either homogeneous or heterogeneous treatment effects--the latter often needed in large population studies. The proposed research will be illustrated using large-scale population health survey data to evaluate the impact of the implementation of the Affordable Care Act on individual health outcomes.

e-mail: *snm@stat.osu.edu*

BAYESIAN INFERENCE ABOUT CAUSAL EFFECTS IN THE PRESENCE OF UNMEASURED CONFOUNDING

Joseph W. Hogan*, Brown University

Allison K. DeLong, Brown University

Michael J. Daniels, University of Texas, Austin

Causal effects from observational studies cannot be identified without relying untestable assumptions about unmeasured confounding. In this talk we describe a Bayesian formulation of a structural causal model that encodes untestable assumptions about unmeasured confounding using a prior distribution. The parameters of this prior are completely unidentifiable in a Bayesian sense: they are independent of observed data, conditionally on other parameters in the model. As a consequence, "sensitivity analysis" about the effects of unmeasured confounding is represented in terms of specification of the priors, which reflects both substance and uncertainty about the untestable assumptions and, perhaps more importantly, are not influenced by the choice of observed-data model. We show that familiar frequentist approaches (e.g., bounds, sensitivity plots) can be reproduced through specific formulations of the prior, providing some insight about the frequentist procedures and their underlying assumptions about unmeasured confounding. We show how to implement the method for both binary and continuous outcomes, and illustrate using data from a recent study of the effect of a government cash transfer program on health outcomes for orphaned and separated children in Kenya.

e-mail: *jhogan@stat.brown.edu*

A BAYESIAN NONPARAMETRIC APPROACH TO MARGINAL STRUCTURAL MODELS FOR POINT TREATMENTS AND A CONTINUOUS OR SURVIVAL OUTCOME

Jason Roy*, University of Pennsylvania

Kirsten Lum, University of Pennsylvania

Michael J. Daniels, University of Texas, Austin

Marginal structural models (MSM) are a general class of causal models for specifying the average effect of treatment on an outcome. These models that can accommodate discrete or continuous treatment, as well as treatment effect heterogeneity (causal effect modification). The literature on estimation of MSM parameters have been dominated by semiparametric estimation methods, such as inverse probability of treatment weighted (IPTW). Likelihood-based methods have received little development, probably in part due to the need to integrate out confounders from the likelihood and due to reluctance to make parametric modeling assumptions. In this paper we develop a fully Bayesian MSM for continuous and survival outcomes that maintains much of the flexibility of semiparametric methods while delivering joint posterior distributions of the causal parameters. We take a Bayesian nonparametric approach, using a combination of a dependent Dirichlet process for the outcome distribution and Gaussian process for the mean to model the observed data. The performance of the methodology is evaluated in several simulation studies.

e-mail: *jaroy@upenn.edu*

98. INTEGRATIVE ANALYSIS OF MULTI-OMIC DATA FOR UNDERSTANDING COMPLEX HUMAN DISEASES

LONGITUDINAL GAUSSIAN GRAPHICAL MODELS INTEGRATE GENE EXPRESSION AND SEQUENCING DATA FOR AUTISM RISK GENE DETECTION

Kevin Lin, Carnegie Mellon University

Han Liu, Princeton University

Kathryn Roeder*, Carnegie Mellon University

Detection of genes influencing autism spectrum disorder (ASD) has been challenging since hundreds of risk genes affect neurodevelopment. Our approach to finding ASD genes overcomes this challenge by integrating brain gene expression data with DNA sequence data. To do so we first estimate a gene interaction network based on co-expression patterns. By assuming that genes interacting with bona fide ASD genes are themselves likely to affect ASD risk, we substantially augment the genetic signal. We build upon an existing technique, called DAWN, which was recently used to analyze network data from BrainSpan samples restricted to a narrow developmental window. This window was chosen to match a key developmental time and region of the brain associated with autism, but the number of samples available is extremely limited. We extend this method by 1) using a novel longitudinal transformation to incorporate more spatio-temporal regions of the BrainSpan data for our gene network estimated by neighborhood selection and 2) using modern high-dimensional hypothesis testing on every edge of the Gaussian graphical model to prune our gene network. By improving the gene network estimation, we detect 200+ genes, roughly a fourth which are validated by new sequencing data - an improvement over previous analyses.

e-mail: r*oeder@stat.cmu.edu*

PRIORITIZATION OF DISEASE-CAUSING GENETIC VARIANTS THROUGH INTEGRATED ANALYSIS OF ASSOCIATION SIGNALS AND GENOMIC FUNCTIONAL ANNOTATION

Qiongshi Lu, Yale University

Ryan Powles, Yale University

Xinwei Yao, Yale University

Yiming Hu, Yale University
Jiehuan Sun, Yale University
Yuwei Cheng, Yale University
Kei Cheung, Yale University
Qian Wang, Yale University
Beixin He, Yale University
Hongyu Zhao*, Yale University

Genome-wide association studies (GWAS) have been a great success in the past decade. However, significant challenges still remain in both identifying new risk loci and interpreting results, even for samples with tens of thousands of subjects. Complex structure of linkage disequilibrium also makes it challenging to separate causal variants from nonfunctional ones in large haplotype blocks. In this presentation, we describe our recent efforts to integrate genomic functional annotations from computational predictions (e.g. genomic conservation) and high-throughput experiments (e.g. the ENCODE and Roadmap Epigenomics Projects) with GWAS test statistics. The effectiveness of our methods will be demonstrated through their applications to several large GWASs. At the single nucleotide polymorphism (SNP) level, top ranked SNPs after prioritization have both higher replication rates and consistently stronger enrichment of eQTLs. Within each risk locus, our methods are also able to distinguish functional sites from groups of correlated SNPs. This is joint work with Qiongshi Lu, Ryan Powels, Xinwei Yao, Yiming Hu, Jiehuan Sun, Yuwei Cheng, Kei-Hoi Cheung, Qian Wang, and Beixin He.

e-mail: *hongyu.zhao@yale.edu*

INTEGRATED ANALYSIS OF DNA METHYLATION, GENETIC VARIATION, AND GENE EXPRESSION DATA IN HUMAN AGING

Karen N. Conneely*, Emory University School of Medicine
Elizabeth M. Kennedy, Emory University School of Medicine
Lynn M. Almli, Emory University School of Medicine
Alicia K. Smith, Emory University School of Medicine
Elisabeth B. Binder, Max Planck Institute of Psychiatry and Emory University School of Medicine
Kerry J. Ressler, Harvard Medical School and Emory University School of Medicine

Epigenome-wide association studies in humans have reported thousands of age-differentially-methylated CpG sites, and recent studies show that age can be predicted from DNA methylation data with great accuracy across a wide range of cell and tissue types. However, the role of these DNA methylation changes remains unelucidated. In whole blood, the profile of associations between age and gene expression is dwarfed by the age-methylation association profile, suggesting that many of the methylation changes observed in whole blood are not directly functional, but may be marks or side effects of another process. In this work, we generate predictions for epigenomic and genomic data under competing mediation and evolutionary models that can potentially explain the observed relationships between DNA methylation and age. We test these predictions as a series of simple tests of enrichment for meQTLs and methylation-expression associations in integrated genomic and epigenomic data from a cross-section of whole blood samples. Our ultimate goal is to explain the widely observed pattern of age-changes in methylation. Though this goal may be best achieved through analysis of longitudinal and/or familial data in multiple tissues, the aim of our current work is to see how far we can get with available cross-sectional microarray data, and to generate preliminary data and predictions for future studies.

e-mail: *kconnee@emory.edu*

INTEGRATIVE MULTI-OMIC ANALYSIS OF X CHROMOSOME INACTIVATION IN EPITHELIAL OVARIAN CANCER

Nicholas B. Larson*, Mayo Clinic
Stacey J. Winham, Mayo Clinic
Zach Fogarty, Mayo Clinic
Melissa C. Larson, Mayo Clinic
Brooke L. Fridley, University of Kansas Medical Center
Ellen L. Goode, Mayo Clinic

In females, X-chromosome inactivation (XCI) epigenetically silences one of the homologous chromosomes, with some genes escaping XCI. This process is tissue- and cell-specific, and the role of XCI in tumors is largely unknown. Using DNA genotypes, RNA-Seq gene expression, and DNA methylation data, we aimed to fully characterize XCI patterns for 113 epithelial ovarian cancer (EOC) patient tumor samples. We first generated allele-specific expression (ASE) for 453 X genes by integrating RNA-Seq with phased genome-wide SNP data. We then identified 89 skewed XCI samples informative for gene-level escapee status using a composite likelihood ratio test and assessed genic XCI on a per sample basis via a two-component beta-binomial mixture model. Although trends of genic XCI escapee patterns were generally concordant with previous LCL studies, we also identified widespread XCI heterogeneity per gene across samples. Additional analyses of available paired tumor-normal data demonstrated somatic alterations of genic XCI status,

indicating potential epigenetic remodeling in EOC. Finally, we investigated predictive models using ASE and methylation to characterize XCI patterns across the entire chromosome and evaluate potential clinical associations. Our findings demonstrate a multi-omic strategy for evaluating XCI in EOC and further research with paired tumor-normal data is necessary.

e-mail: *larson.nicholas@mayo.edu*

99. ANALYSIS OF NEXT-GENERATION SEQUENCING DATA: INCREASING ACCURACY AND NOVEL APPLICATIONS

THE NEXT GENERATION OF (EPI-) GENOMIC DATA: SINGLE CELLS

Faye Zheng, Purdue University

Rebecca W. Doerge*, Purdue University

Within the last decade sequencing technologies have scaled rapidly with respect to both throughput and accuracy. Independent of the material being sequenced (DNA or RNA), the capability of these new (next-generation) technologies to profile the molecular content of individual cells is now reality. Since the behavior of cells is dynamic, and conditioned on the setting in which the cell(s) exist (e.g., cancer, differentiation, etc.), the scientific community has great interest in interrogating cell-to-cell heterogeneity and understanding its biological consequences. At this point, the analysis of single-cell data is largely exploratory and significantly lacking from valid statistical investigations. As such, there is an opportunity for the analysis of single cell genomic and epigenomic data to benefit from proper experimental design, predictive modeling and hypothesis-based testing. The statistical and technological issues of single cell sequencing data will be discussed. Both simulated and real data will be employed to illustrate the unique features of single cell data as compared to current bulk-cell data. The hope for this talk is that it will stimulate awareness and thought from the statistical community on how to address data management, analysis and interpretation of these large complex data.

e-mail: *doerge@purdue.edu*

DISPENSING WITH THE BIOINFORMATICS PIPELINE: STATE SPACE MODELS FOR NGS BASE-CALLING AND ERROR CORRECTION

Karin S. Dorman*, Iowa State University

Xin Yin, Iowa State University

Aditya Ramamoorthy, Iowa State University

The age-old wisdom “garbage in, garbage out” underscores any analysis using next-generation sequencing (NGS) data. NGS errors are typically mitigated through two consecutive stages in the data analysis pipeline, base-calling followed by error-correction. Interestingly, base callers rarely utilize information available about the underlying genome sequence, whereas error-corrections methods seldom utilize properties of the sequencing machine. We have demonstrated that a probabilistic approach to error correction beats all existing algorithmic methods. Even better, an integrated approach that combines the information available to both base-calling and error-correction methods improves performance over their serial application. Menges et al. (2011) has previously proposed a base caller that borrows information from alignment to a known reference genome. In contrast, our goal is to utilize genome information in the absence of a known reference genome. Specifically, we use a Hidden Markov Model, where the transition distribution captures local genomic dependence, and the emission distribution models intensities. The combined method removes at least twice as many errors Bustard, the standard Illumina base caller.

e-mail: *kdorman@iastate.edu*

CAN EPIGENETIC PROFILES PREDICT GENETIC RISK FOR BLOOD DISORDERS?

Paul Auer*, University of Wisconsin, Milwaukee

Alex Reiner, Fred Hutchinson Cancer Research Center

Genome-wide association studies (GWAS) of common genetic variants have implicated 68 independent loci that underlie circulating platelet counts (PLT). Because these loci explain a small portion of the overall genetic contribution, recent attention has focused on the role of rare genetic variants. However, it is well-known that rare-variant association studies are underpowered to detect modest effect sizes for complex traits such as PLT. In order to enhance the power to detect associations, we sought to utilize epigenetic data from the RoadMap Epigenomes project to prioritize genomic regions for analysis. Using RoadMap data from 127 cell-types, we implemented a number of different machine learning approaches for predicting the impact of genetic variants on PLT. These models were trained on common variant GWAS results, and tested on imputed whole-genome sequencing data on approximately 20,000 individuals. Simulations and real data analysis demonstrate the usefulness of using epigenetic data to inform genetic association studies.

e-mail: *pauer@uwm.edu*

DECONVOLVING COPY NUMBER PROFILES OF CANCER GENOMES USING NGS DATA

Xuefeng Wang*, Stony Brook University

Genome-wide DNA copy number aberrations (CNA) analysis of large numbers of tumor samples is a vital step to identify cancer

driver events and gain insights into the cancer progression and prevention. The first part of this talk will introduce a set of concepts and analytic tools that enables the fast and accurate dissection of CNA profiles by integrating the totality of information available from NGS data. Based on a large scale whole exome sequencing analysis of melanoma samples, we demonstrate that the described approaches allow more accurate estimation of copy number status and the identification of otherwise undetectable CNA patterns both across sites and across samples. In the second part, we present novel methods to characterize tumor cell populations based on gene copy number signals obtained from low-coverage single-cell DNA sequencing. A set of segmentation approaches are introduced for systematic breakpoint fine-location. We show that the proposed strategy provides effective denoising of low-coverage single cell DNA-seq data and allow more exact tumor profiling. We demonstrate our methods via simulations and apply them to a prostate cancer data.

e-mail: *pwxfor@gmail.com*

100. NETWORKS FOR HIGH DIMENSIONAL TIME SERIES

LAG STRUCTURED MODELING FOR HIGH DIMENSIONAL VECTOR AUTOREGRESSION

William Nicholson, Cornell University

Jacob Bien*, Cornell University

David Matteson, Cornell University

Vector autoregression (VAR) is a fundamental tool for modeling the joint dynamics of multivariate time series. However, as the number of component series is increased, the VAR model quickly becomes overparameterized, making reliable estimation difficult in high dimensional settings. A common assumption in time series is that the dynamic dependence among components is short-range, leading to the common practice of lag order selection. We propose a new class of regularized VAR models that embeds the notion of lag selection into a convex regularizer. The key convex modeling tool is a group lasso with hierarchically nested groups that guarantees that the sparsity pattern of autoregressive lag coefficients honors the ordered structure inherent to VAR. We provide computationally efficient algorithms for solving this problem and demonstrate improved performance in forecasting and lag order selection over previous approaches. A macroeconomic application highlights the convenient, interpretable output of our method.

e-mail: *jbien@cornell.edu*

NON-GAUSSIAN ESTIMATION FOR TIME SERIES SAMPLED AT MIXED FREQUENCIES

Alex Tank, University of Washington

Emily Fox*, University of Washington

Ali Shojaie, University of Washington

Time series sampled at dierent sampling rates are common in econometrics. Most models and methods for this type of data are based on a joint Gaussian assumption of the error residuals. Unfortunately, this assumption leads to non-identifiability for an underlying VAR model that evolves at the rate of the finest sampled series. Recently, a method has been proposed to infer underlying VAR models for subsampled time series based on a non-Gaussian assumption of the error residuals leading to a fully identifiable model. This method was developed only for time series where the subsampling is the same across time series. We present an extension bringing this non-Gaussian methodology to allow model identifiability for time series sampled at different sampling rates.

e-mail: *ebfox@uw.edu*

NETWORK RECONSTRUCTION FROM HIGH DIMENSIONAL ORDINARY DIFFERENTIAL EQUATIONS

Shizhe Chen, University of Washington

Ali Shojaie*, University of Washington

Daniela Witten, University of Washington

We consider the task of learning a dynamical system from high-dimensional time-course data, for instance, for estimating a gene regulatory network from gene expression data measured at discrete time points. We model the dynamical system non-parametrically as a system of additive ordinary differential equations. Most existing methods for parameter estimation in ordinary differential equations estimate the derivatives from noisy observations. This has been shown to be challenging and inefficient. We propose a novel approach that does not involve derivative estimation. We show that the proposed method can consistently recover the true network structure even in high dimensions, and we demonstrate empirical improvement over competing approaches.

e-mail: *ashojaie@uw.edu*

101. ENVIRONMENTAL AND ECOLOGICAL APPLICATIONS

AN ADAPTIVE ASSOCIATION TEST FOR MICROBIOME DATA

Chong Wu*, University of Minnesota

Junghi Kim, University of Minnesota

Wei Pan, University of Minnesota

Distance-based analysis and microbiome regression based kernel association test (MiRKAT) are two popular methods for investigating how the compositions of microbial communities are associated with various risk factors. A proper choice of a phylogenetic distance is critical for the power of these methods. However, existing phylogenetic distance metrics are designed without accounting for differential information contents with various microbial lineages, leading to power loss in the association testing. We propose a class of microbiome based sum of powered score (MiSPU) tests based on a newly defined generalized taxon proportion that combines observed microbial composition with phylogenetic tree information. Different from the existing methods, a MiSPU test is based on a weighted score of the generalized taxon proportion, upweighting more likely to be associated microbial lineages. Our simulations demonstrated that one or more MiSPU tests were more powerful than MiRKAT while correctly controlling type I error rates. An adaptive MiSPU (aMiSPU) test is proposed to combine multiple MiSPU tests with various weights, approximating the most powerful MiSPU for a given scenario, consequently being highly adaptive and high powered across various scenarios. In the end, we applied MiSPU and aMiSPU to a throat microbiome dataset.

e-mail: *wuxx0845@umn.edu*

A CLASS OF DISTANCE TESTS FOR COMPARING ENVIRONMENTAL EXPOSURE DISTRIBUTIONS IN PRESENCE OF DETECTION LIMITS

Yuchen Yang*, University of Kentucky

Brent Shelton, University of Kentucky

Richard Kryscio, University of Kentucky

Tom Tucker, University of Kentucky

Li Li, Case Western Reserve University

Li Chen, University of Kentucky

In environmental exposure studies, a fundamental question of interest is to compare chemical exposure distributions between two groups. One complexity for addressing this question is that a portion of exposure measurements may fall below experimentally determined detection limits (DLs), which results in left-censored data. To analyze such data, nonparametric methods have received increasing attentions recently because of their robustness. Current nonparametric methods, including the reverse log-rank test and Peto-Peto test, transform ("flip") data subject to DLs to right-censored and then apply log-rank test and Peto-Peto test for right-censored data to make inferences. However, these methods are not epidemiologically meaningful because they are constructed by comparing hazard functions essentially, which are not meaningful quantities for environmental exposure data. In addition, they are inefficient because they are rank-based and not sensitive to the magnitude of difference between two distributions. Furthermore, these methods can yield to inflated type I errors when the DL distributions in the two groups differ. We propose a class of nonparametric test statistics by considering the integrated weighted difference in kernel-based estimators for the cumulative distribution functions of the two groups. Simulation studies demonstrate that the proposed tests preserve type I errors regardless whether the DL distributions in the two groups differ or not and are more efficient than current methods in certain situations. An analysis of a colon cancer study is provided for illustration.

e-mail: *yuchen.y@uky.edu*

DIETARY PATTERNS AND DETERMINANTS OF MERCURY AND OMEGA-3 EXPOSURE IN PREGNANT WOMEN LIVING IN THE SEYCHELLES

Tanzy Love*, University of Rochester

Maria Mulhern, Ulster University

Sally Thurston, University of Rochester

Alison Yeates, Ulster University

Katie Evans, Dupont

Maxine Bonham, Ulster University

Emeir McSorley, Ulster University

Conrad F. Shamlaye, Seychelles

J. J. Strain, Ulster University

P**hilip W. Davidson**, University of Rochester

Current fish advisories are based on epidemiological studies that associate fetal exposure to methylmercury (MeHg), presumably from fish consumption, with the children's developmental outcomes. However, the relationship between fetal exposure and fish consumption is not straightforward. There are possible interactions with nutrients present in fish such as omega-3 long chain poly unsaturated fatty acids (LC-PUFA) and with foods eaten as part of a dietary pattern. We examined dietary patterns to determine the overall dietary associations. Dietary patterns represent a broader picture of nutrient consumption and may be more relevant for fetal development than individual foods or nutrients. We compared fish consumption and dietary patterns to discover determinants of MeHg and LC-PUFA status in pregnant women. Cluster analysis was used to determine dietary patterns, and we correlated these dietary patterns with MeHg and LC-PUFA biomarkers. This work further explores the complex interrelationships present between nutrition and possible adverse associations with prenatal MeHg exposure in the Seychelles longitudinal observational cohort study.

e-mail: *tanzy_love@urmc.rochester.edu*

SPATIAL PREDICTION OF NATURALLY OCCURRING INDOOR GAMMA RADIATION IN GREAT BRITAIN

Pavel Chernyavskiy, National Cancer Institute, National Institutes of Health

Gerald M. Kendall, University of Oxford

Philip S. Rosenberg, National Cancer Institute, National Institutes of Health

Richard Wakeford, University of Manchester

Mark P. Little*, National Cancer Institute, National Institutes of Health

Gamma radiation from natural sources is an important component of background radiation, and correlates with childhood leukemia risk in Great Britain. The geographic variation of indoor gamma radiation dose-rates in Great Britain is explored using various geostatistical methods. A multi-resolution Gaussian process (MRGP) model using radial basis functions with 2, 4, or 8 components, is fitted via maximum likelihood, and a non-spatial model is also used, fitted by ordinary least squares; because of the dataset size (N=10,199), four other parametric spatial models are fitted by variogram-fitting. A randomly selected 70:30 split is used for fitting:validation. The models are evaluated based on their predictive performance as measured by Mean Absolute Error, Mean Squared Error, as well as Pearson correlation and rank-correlation between predicted and actual dose-rates. Each of the four parametric models (Matérn, Gaussian, Bessel, Spherical) fitted the empirical variogram well, and yield similar predictions at >50 km separation, although with more substantial differences in predicted variograms at <50 km. The MRGP model with 8 components had the best predictive accuracy among the models considered. The Spherical, Bessel, Matérn, Gaussian and ordinary least squares models had progressively worse predictive performance, the ordinary least squares model being particularly poor in this respect.

e-mail: *mark.little@nih.gov*

BAYESIAN DISTRIBUTED LAG INTERACTION MODELS

Ander Wilson*, Harvard School of Public Health

Yueh-Hsiu Mathilda Chiu, Icahn School of Medicine at Mount Sinai

Rosalind Wright, Icahn School of Medicine at Mount Sinai

Brent Coull, Harvard School of Public Health

A growing body of research supports an association between maternal exposure to air pollution during pregnancy and children's health. Recent research has focused on estimating critical windows of vulnerability and estimating exposure effect heterogeneity. Simultaneous estimation of windows of vulnerability and effect heterogeneity is typically accomplished by fitting a distributed lag model (DLM) stratified by subgroup. However, this does not allow for subgroups to have the same window of vulnerability but different effects within the window or to have different windows but the same within-window effect. We propose a new approach that partitions the DLM into a constrained functional predictor that estimates windows of vulnerability and a scalar effect size that estimates the effect within the window. By allowing either component to be shared or vary across group, the proposed method allows for heterogeneity in only the window of vulnerability, only the effect within the window, or in both. In a simulation study we show that sharing a component across groups results in improved estimation of the windows and effect. We estimate the effect of prenatal exposures to fine particulate matter on birth weight and asthma in a prospective cohort study and identifying heterogeneity by sex and maternal obesity status.

e-mail: anderwilson@gmail.com

ESTIMATION OF THE EFFECTIVENESS OF INFLUENZA VACCINATION FROM HOUSEHOLD STUDIES

Kylie Ainslie*, Emory University

Michael Haber, Emory University

Influenza vaccination is now recommended in the U.S. to all individuals greater than six months of age, making randomized clinical trials unethical. Therefore, vaccine effectiveness (VE) is estimated yearly using observational studies. Recent studies have followed households rather than separate individuals to determine VE against influenza transmission from the household compared to VE against influenza transmission from the community. Here, we present a likelihood approach to estimate vaccine-related protection against transmission of influenza from the household and from the community when the source of infection (household or community) is known. We use symptomatic influenza as our outcome of interest and allow for vaccination to occur at any time within the study period. Previous methods have required vaccination to occur only at the beginning of the study. An additional advantage of our approach is that maximum likelihood estimation does not rely on the assumptions required by the proportional hazards model. We then extend our likelihood approach to the scenario when infection source is unknown. Finally, we further extend our approach to estimate strain-specific VE.

e-mail: *kylie.ainslie@emory.edu*

A STOCHASTIC EPIDEMIC MODEL TO IDENTIFY UNOBSERVED URBAN INSECT INFESTATIONS

Erica M.W. Billig*, University of Pennsylvania

Jason A. Roy, University of Pennsylvania

Michael Z. Levy, University of Pennsylvania

The analysis of epidemic dynamics are complicated by several factors, including the fact that the true dispersal mechanism of disease agents and the true infection times of patients are typically unobserved. Instead, we often observe the infection state of a subset of individuals at a specific time, which causes many unobserved variables. The likelihood of the model quickly becomes intractable given these unknowns. We propose a novel stochastic compartmental model to analyze urban insect infestations that incorporates the counts of disease vectors at each house and complex spatial dispersal dynamics. We apply our method to a recent study of the Chagas disease vector Triatoma infestansin Arequipa, Peru. Our goal of the analysis is to predict and identify houses that are infested with T. infestans. The data are obtained two ways: an inspector is sent out to pro-actively inspect houses, or residents reactively report infestations, which are then verified by an inspector. The Bayesian method is used to augment the data, estimate the insect population growth and dispersal parameters, and determine posterior infestation risk probabilities of households. We investigate the properties of the model with simulation studies and analyze the Chagas disease vector data to create an informed control strategy.

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

102. GENOMICS

DETECTING eQTLs IN MEGAKARYOCYTES (MKS) DERIVED FROM INDUCED PLURIPOTENT STEM CELLS (iPSCs)

Kai Kammers*, Johns Hopkins Bloomberg School of Public Health

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

Ingo Ruczinski, Johns Hopkins Bloomberg School of Public Health

Joshua Martin, The GeneSTAR Program, Johns Hopkins School of Medicine

Margaret A. Taub, Johns Hopkins Bloomberg School of Public Health

Lisa R Yanek, The GeneSTAR Program, Johns Hopkins School of Medicine

Linzhao Cheng, Johns Hopkins School of Medicine

Zack Z. Wang, Johns Hopkins School of Medicine

Rasika A. Mathias, The GeneSTAR Program, Johns Hopkins School of Medicine

Lewis Becker, The GeneSTAR Program, Johns Hopkins School of Medicine

Aggregation of platelets in the blood may initiate arterial occlusions causing heart attacks or strokes. To understand the biology of platelet aggregation, we examined genetic and transcriptomic data from megakaryocytes (MKs), the precursor cells for anucleate platelets, that are derived from induced pluripotent stem cells (iPSCs). Our goal is to detect patterns of transcript expression in the MKs related to specific genetic variants. In this presentation we show our analysis pipeline for performing extensive eQTL analyses - from raw RNA-seq reads and genotype data to eQTL plots showing gene-SNP interactions. We explain in detail how transcript expression data are filtered, transformed, and batch corrected. We also discuss possible pitfalls and artifacts that may occur when analyzing genomic data from different sources jointly. To date, our data contains MKs from 161 people with informative genotypes. Given a high genetic and transcriptomic integrity of the iPSC-derived MKs, we found several hundred cis-eQTLs in European Americans and African Americans and see a high replication between the two groups. The majority of cis-eQTLs are unique to MKs compared to other tissue types that are reported in GTEx Portal.

e-mail: *kai.kammers@jhu.edu*

DETECTING RARE HAPLOTYPE-ENVIRONMENT INTERACTION UNDER UNCERTAINTY OF GENE-ENVIRONMENT INDEPENDENCE ASSUMPTION

Yuan Zhang*, University of Texas, Dallas

Shili Lin, The Ohio State University

Swati Biswas, University of Texas, Dallas

Finding rare variants and gene-environment interactions (GXE) is critical in the quest for “missing heritability” in complex diseases. We consider the challenging problem of detecting GXE where G is a rare haplotype variant and E is a non-genetic factor. A common assumption made in such analyses is independence of G and E in the controls or in the target population. As the assumption may not hold in general, developing methods that do not make this assumption yet retains good powers to detect GXE is of important practical interest. To this end, we consider a recently proposed method logistic Bayesian LASSO (LBL) for detecting GXE using case-control sample, which assumes G-E independence, referred to as LBL-GXE-I. It has inflated type I error rates when G-E independence assumption is violated. We propose a way to relax this assumption by modeling the haplotype frequencies as functions of E through a multinomial logistic regression model. We refer to it as LBL-GXE-D and show that it controls type I error rates in all situations. However, LBL-GXE-D has reduced power than LBL-GXE-I when G-E independence holds, as expected. To optimize power without sacrificing type I error in any scenario, we propose a unified approach by employing a reversible jump Markov chain Monte Carlo method. It allows moves between G-E dependence and independence models of different dimensionalities and thus incorporates uncertainty in G-E indepen-

dence assumption — we refer to it as LBL-GXE. Our extensive simulation studies show that LBL-GXE has power similar to that of LBL-GXE-I in case of G-E independence, and at the same time has controlled type I errors in all situations. Finally, we analyze a lung cancer dataset and find significant interaction between a specific rare haplotype in 15q25.1 region and smoking in the presence of G-E dependence.

e-mail: *yxz112020@utdallas.edu*

A NUMERICAL METHOD FOR LIKELIHOOD ESTIMATION OF SPECIES TREES FROM LARGE GENOMIC DATA USING THE COALESCENT PROCESS

Arindam RoyChoudhury*, Columbia University

A species tree is a weighted tree-graph that represents the order and the magnitude of genetic separation between a given set of species. Statistical estimation of trees from genomic data is an integral part of studying species trees. Although various likelihood and Bayesian estimators of species trees are available, none of these methods are fast enough to estimate very large species trees under a certain commonly used model (the coalescent). This problem is especially relevant today because there has been a recent influx of large amount of genomic data. Here I will present an approach of fast likelihood estimation of species trees, computationally exploiting a certain special structure of the tree space. Using this approach, one will be able to numerically estimate species trees from larger genomic data than previously possible in a reasonable time.

e-mail: *ar2946@columbia.edu*

A STATISTICAL FRAMEWORK FOR PREDICTIVE MODELING OF MICROBIOME DATA INTEGRATING THE PHYLOGENETIC TREE

Jun Chen*, Mayo Clinic

Jian Xiao, Mayo Clinic

The human microbiome, which has now been regarded as the "extended" human genome, has attracted considerable attention from both biomedical scientists and clinical investigators. Numerous studies have revealed a significant role of the human microbiome in disease development and prognosis. Thus the human microbiome can be potentially used for precision medicine to improve patient care. Integrating the human microbiome data into medicine requires developing powerful predictive models while taking into the special characteristics of microbiome data. One popular type of microbiome data is generated by sequencing a region of the bacterial 16S rRNA gene. The output of the 16S targeted sequencing is an abundance table of the detected bacterial species, along with their phylogenetic relationship. Utilizing the phylogeny information in microbiome-based prediction is critical since the microbial taxa tend to be associated with the phenotype at various phylogenetic resolutions. However, efficient use of the phylogeny raises some challenges. Here I will present a framework for integrating the phylogenetic tree into predictive modeling of the microbiome data. I will introduce a new type of sparse regression model, inverse-correlation matrix regularized sparse regression (ICM-SR), where the correlation matrix is defined based on the phylogeny. Simulation as well as real 16S data will be used to illustrate the proposed method.

e-mail: *jchen1981@gmail.com*

PROMISE-ME: A ROBUST METHOD FOR INTEGRATED ANALYSIS OF DNA METHYLATION, GENE EXPRESSION, AND MULTIPLE BIOLOGICALLY RELATED CLINICAL AND PHARMACOLOGICAL OUTCOMES

Xueyuan Cao,

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

Tong Lin, St. Jude Children's Research Hospital

Projection onto the Most Interesting Statistical Evidence (PROMISE) is a robust method to perform integrated analysis of one form of genomic data with multiple biologically related pharmacologic and clinical outcome variables. PROMISE has been used successfully to identify genomic and transcriptomic features of pharmacogenetic relevance to the treatment of pediatric leukemia. Here, we extend that framework to develop PROMISE for methylation and expression (PROMISE-ME) as a robust method to perform integrated analysis of DNA methylation, gene expression, and multiple pharmacologic and clinical outcome variables. For each gene, PROMISE-ME evaluates the association of methylation with expression, the association of methylation with each outcome variable, and the association of expression with each outcome variable. Previous knowledge of the biological relationships among outcome variables is used to define a test statistic that is a linear combination of each of the pairwise association statistics described above. Permutation is used to determine the significance of this test statistic. The advantages of PROMISE-ME in terms of enhancing statistical power for meaningful biological discoveries are seen in simulation studies and examples from pediatric oncology research.

e-mail: *stanley.pounds@stjude.org*

INTEGRATED ANALYSIS OF MULTIDIMENSIONAL (EPI) GENETIC DATA ON CUTANEOUS MELANOMA PROGNOSIS

Yu Jiang*, University of Memphis

Xingjie Shi, Nanjing University of Finance and Economics, China

Qing Zhao, Yale University

Shuangge Ma, Yale University

Cutaneous melanoma poses a serious public health concern. Multiple types of genomic changes have been implicated in its prognosis. Many of the existing studies, especially the early ones, are limited in analyzing a single type of genomic measurement and cannot comprehensively describe the biological processes underlying prognosis. As a result, the obtained prognostic models can be less satisfactory, and the identified prognostic markers tend to be less informative. The recently collected TCGA (The Cancer Genome Atlas) cutaneous melanoma data have high-quality, comprehensive genomic measurements, making it possible to more comprehensively and accurately modeling prognosis. In this study, we first describe statistical approaches that are able to integrate multiple types of genomic measurements with the assistance of variable selection and dimension reduction techniques. The analysis of TCGA data suggests that integrating multi-measurements will lead to prognostic models with improved prediction performance. In addition, informative individual markers and pathways are identified, which provide valuable insights into melanoma prognosis potentially.

e-mail: *yjiang4@memphis.edu*

DETECTION OF SHARED COMMON GENETIC VARIANTS BETWEEN COMPLEX DISEASE PAIRS

Julie Kobie*, University of Pennsylvania

Sihai D. Zhao, University of Illinois, Urbana-Champaign

Yun R. Li, The Children's Hospital of Philadelphia

Hakon Hakonarson, The Children's Hospital of Philadelphia

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. This paper investigates 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 single nucleotide polymorphisms (SNPs) associated with complex diseases with increased sample sizes, but standard meta-analysis approaches are not optimal for the study of these diseases. This paper presents 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 that share at least one common genetic variant.

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

103. META-ANALYSIS

INTEGRATIVE ANALYSIS FOR PATHWAY SELECTION USING INDIVIDUAL PATIENT DATA

Quefeng Li*, University of North Carolina, Chapel Hill

Menggang Yu, University of Wisconsin, Madison

Sijian Wang, University of Wisconsin, Madison

Multiple related genetic studies are often analyzed together to identify genes associated with prognosis. In this era of big data, as the individual patient data (IPD) become more accessible, the integrative analysis using IPD are now extensively conducted. The existing integrative analysis methods aim to identify prognostic genes. It has been recognized that genes do not work alone, but in networks or pathways of interactions. Involving the external pathway information into the analysis is more appealing as it improves understanding of the disease process. In this paper, we propose a new integrative analysis method that uses the knowledge of pathways. Our method employs a hierarchical decomposition followed by a proper regularization to identify important pathways across multiple studies. Theories are provided to show that our method can asymptotically correctly identify pathways and their important members (genes). We explicitly show that pathway identification needs milder conditions than gene identification as it allows certain false positives/negatives at the gene selection level. Our method also handles cases when pathways have overlapping genes. Its finite-sample performance is shown to be superior than other ad hoc methods in various simulation studies. We further apply our method to analyze five cardiovascular diseases studies.

e-mail: *quefeng@email.unc.edu*

ALTERNATIVE MEASURES OF BETWEEN-STUDY HETEROGENEITY IN META-ANALYSIS: REDUCING THE IMPACT OF OUTLYING STUDIES

Lifeng Lin*, University of Minnesota

Haitao Chu, University of Minnesota

James S. Hodges, University of Minnesota

Meta-analysis has become a widely used tool to combine results from separate studies. The collected studies are homogeneous if they share a common underlying true effect size; otherwise, they are heterogeneous. A fixed-effects model is customarily used when the studies are homogeneous, while a random-effects model is used for heterogeneous studies. Assessing heterogeneity in meta-analysis is critical for model selection. Ideally, if heterogeneity is present, it should

permeate the entire collection of studies, instead of being limited to a small number of outlying studies. Outliers can have great impact on the conventional measures of heterogeneity and the conclusions of a meta-analysis. However, no widely accepted guidelines exist for handling outliers. This article proposes several new heterogeneity measures. In the absence of outliers, the proposed measures are close to the conventional ones; in the presence of outliers, the proposed measures are less affected than the conventional ones. The performance of the proposed and conventional heterogeneity measures are compared theoretically, by studying their asymptotic properties, and empirically, using simulations and case studies.

e-mail: *linl@umn.edu*

TESTING FOR PUBLICATION BIAS UNDER THE COPAS SELECTION MODEL IN META-ANALYSIS

Yong Chen, University of Pennsylvania

Jing Ning, University of Texas MD Anderson Cancer Center

Jin Piao*, University of Texas Health Science Center, Houston

In meta-analyses, publication bias is a well-known, important and challenging issue because if the sample of studies retrieved for review is biased, the validity of the results of a meta-analysis is threatened. One of popular methods is the Copas selection model, which is a flexible framework for correcting the estimates with considerable insight into the publication bias mechanism. However, rigorous test procedures under the Copas selection model to detect the bias are lacking. To fill the gap, we develop a score based test for detecting publication bias under the Copas selection model. The statistical challenge is that the asymptotic behavior of the proposed test statistic is irregular because some parameters disappear under the null hypothesis, leading to identifiability problem. We conduct extensive Monte Carlo simulations to evaluate the performance of the proposed test and illustrate the method using a real data example.

e-mail: *jin.piao@uth.tmc.edu*

PARAMETRIC BOOTSTRAP TO CONSTRUCT CONFIDENCE INTERVALS FOR EVENT RATES OR DIFFERENCES IN RATES IN META-ANALYSES

Gaohong Dong*, Novartis Pharmaceuticals Corporation

Roland Fisch, Novartis Pharma AG

Jennifer Ng, Novartis Pharmaceuticals Corporation

Steffen Ballerstedt, Novartis Pharma AG

Marc Vandemeulebroecke, Novartis Pharma AG

WITHDRAWN

The paediatric investigational plan for Certican® (everolimus) included paediatric studies in liver and kidney transplantation. However, the very slow enrolment made it impossible to recruit patients in a timely manner. Following the recent EMA concept paper on extrapolation (EMA, 2013) and with consultations with EMA, we developed an extrapolation methodology bridging adult and paediatric data via meta-analyses, which included 57 adult studies (>19500 patients) and 7 paediatric studies (651 patients). For the extrapolation, the meta-analytic model is parameterized such that the covariate effects are linear on the log odds scale. For the estimated event rates and differences in rates, initially we obtained confidence intervals (CIs) based on the delta method directly from PROC NLMIXED. However these intervals were deemed unsatisfactory, since the nonlinearity of the parameter transformation led to poor coverage properties and to nonsensical CI limits (i.e. negative lower limits). Therefore, we derived CIs via a parametric bootstrap approach. In this talk, we will briefly introduce maximum likelihood and Bayesian meta-analysis models, then focus on the parametric bootstrap to construct CIs for event rates or differences in rates. We will also explain why this bootstrap approach is needed for maximum likelihood meta-analyses, but not for Bayesian meta-analyses.

e-mail: *gaohong.dong@novartis.com*

USING META-ANALYTIC APPROACHES FOR ANALYZING NON-CONVERGING CLUSTERED DEPENDENT BINARY DATA

Aobo Wang*, Virginia Commonwealth University

Roy T. Sabo, Virginia Commonwealth University

Clustered-correlated data often feature nested random effects and repeated measures. If coupled with binary outcomes and large samples (>10,000), this complexity can lead to non-convergence problems for the desired model. To work around this problem we split data intokindependent, mutually-exclusive sub-samples, or use the natural existing clusters as subsamples, to fit models that converged while accounting for the desired cluster structure. Estimates from these subsamples were then recombined using random-effect-based meta-analytic approaches. We present simulation studies comparing the performance of using the smallest possible number of independent subsamples versus using natural existing clusters, and investigate both the incorporation and exclusion of possible dependence between samples. These studies show that using natural clusters leads to larger inter-cluster variances and standard errors of the requisite test statistic and lower power for the corresponding test, as compared to the independent sample approach, while the inter-cluster variances and standard errors produced by the independent sample approach increased with the number of sub-samples, as expected. We also apply these methods to data on cancer screening behaviors obtained from electronic health records from n=15,652 individuals. These results support the conclusions from the simulation studies, showing that the independent sample approach provides

estimates that are closer to those from the full dataset than does the cluster based approach.
e-mail: w*anga3@vcu.edu*

A LINEUP PROTOCOL FOR FUNNEL PLOT ASSESSMENT IN META-ANALYSIS

Michael P. LaValley*, Boston University School of Public Health

Meta-analyses of studies identified through searching the published literature may be affected by publication bias, the tendency for preferential publication of statistically significant results. This creates a need to assess meta-analytic results for bias. While statistical tests to detect bias are available, their power is often low for meta-analyses with a limited number of studies. As a consequence, the visual inspection of symmetry in funnel plots is often used for bias assessment, but the interpretation of these plots can be subjective. The lineup protocol, named after the lineup of suspects in police investigations, has been proposed as way to make valid statistical inferences from plots for the purposes of exploratory data analysis and model diagnostics. In this protocol, the assessor is asked to select the real data plot from among a set of plots generated under the null hypothesis to prevent over-interpretation of chance variations in the data. To provide more valid judgements of bias I develop a lineup protocol for visual assessment of symmetry in funnel plots from meta-analytic studies. The use of this protocol for funnel plots will be demonstrated with several examples.
e-mail: *mlava@bu.edu*

104. SEMI-PARAMETRIC AND NON-PARAMETRIC METHODS

SINGLE INDEX MODELING AND ESTIMATION IN SECONDARY ANALYSIS OF CASE-CONTROL STUDIES

Liang Liang*, Texas A&M University

Raymond J. Carroll, Texas A&M University

Yanyuan Ma, University of South Carolina

Studying the relationship between covariates based on retrospective data is the main purpose of secondary analysis, an area of increasing interest. We examine the secondary analysis problem when multiple covariates are available, while only a regression mean model is specified. Despite the completely parametric modeling of the regression mean function, the case-control nature of the data requires special treatment and semiparametric efficient estimation generates various nonparametric estimation problems with multivariate covariates. We devise a single index approach that fits with the specified primary and secondary models in the original problem setting, and use reweighting to adjust for the case-control nature of the data, even when the disease rate in the source population is unknown. The resulting estimator is both locally efficient and robust against the misspecification of the regression error distribution, which can be heteroscedastic as well as non-Gaussian. We demonstrate the advantage of our method over several existing methods both analytically and numerically.
e-mail: *lliang021990@gmail.com*

MARGINAL MEAN MODELS FOR ZERO-INFLATED COUNT DATA WITH SPLINE-BASED SEMIPARAMETRIC ESTIMATION

David Todem*, Michigan State University

Yifan Yang, Michigan State University

Wei-Wen Hsu, Kansas State University

KyungMann Kim, University of Wisconsin, Madison

We propose a semiparametric zero-inflated regression model for count data that directly relates covariates to the marginal mean response representing the desired target of inference. The model specifically assumes two semiparametric forms for the log-linear model for the marginal mean and the logistic-linear model for the susceptible probability, in which the fully linear models are replaced with partially linear link functions. A spline-based estimation is proposed for the nonparametric components of the model. Asymptotic properties for the estimators of the parametric and the nonparametric components of the models are discussed. Specifically, we show that the estimators are shown to be consistent and asymptotically efficient under mild regularity conditions. Simulation studies are conducted to evaluate the finite sample performances of this estimation method. Finally, the model is applied to dental caries indices in low income African-American children to evaluate the effects of sugar intake on caries development, taking into account important confounders such as age.
e-mail: *todem@msu.edu*

NONPARAMETRICALLY ASSISTED PARAMETRIC REGRESSION ANALYSIS FOR MULTIPLE-INFECTION GROUP TESTING DATA

Dewei Wang*, University of South Carolina

Peijie Hou, University of South Carolina

WITHDRAWN

Group testing has been recognized as an efficient tool in saving testing expenditures for large-scale screening practices. Typically it has been used for binary disease screening. Recently, the use of multiplex assays has transformed its goal to detecting multiple infections simultaneously. The main goal of this project is to develop parametric regression methods for each infection using multiple-infection group testing data, where the potential correlation among unobserved individual disease statuses cannot be ignored. Previous techniques assume the correlation does not depend on

individual level information and can only be applied to data arising from "master" pool testing. Specifically, these techniques are not designed to make use of retesting information gained from decoding pools that are initially tested positive for at least one infection. In this talk, I will introduce a new method that can provide valid statistical inference for each infection, can allow the correlation to be dependent on individual level information, and can be applied to all types of multiple-infection group testing data. We provide the asymptotic properties of our estimator. Further we illustrate the finite sample performance of our method through simulation and by applying it to chlamydia and gonorrhea screening data collected from the Infertility Prevention Project.

e-mail: *deweiwang@stat.sc.edu*

NONPARANORMAL GRAPHICAL MODEL ESTIMATION WITH FALSE DISCOVERY RATE CONTROL: A SCORE TEST APPROACH

Ritwik Mitra*, Princeton University

Yang Ning, Princeton University

Han Liu, Princeton University

Estimation of graphical model structures for semi-parametric family of distributions have received a lot of attention in recent days. The usual estimation is performed via regularization under sparsity assumptions. In this paper, we study estimation of nonparanormal graphical models via a multiple testing approach. We estimate the edges of the graph and then perform multiple testing in order to control the false discovery rate. The testing for the individual edges in the multiple testing is carried out via score tests.We show that this method asymptotically controls the false discovery rate in detecting the presence of edges. We perform empirical studies on synthetic as well as real data to show the efficacy of this methodology.

e-mail: *rmitra@princeton.edu*

TESTING FOR ASSOCIATION IN A HETEROGENEOUS SAMPLE

Fangyuan Zhang*, Texas Tech University

Jie Ding, Stanford University

Shili Lin, The Ohio State University

It is challenging to identify relationships between variables when only a subset is correlated, and the situation becomes even harder when subsets are correlated in different directions. By ranking the relationship according to Kendall's tau, a tau-path can be derived to facilitate the identification of correlated subsets, if such exits. However, current algorithms for finding the tau-path may only achieve suboptimum given its greedy nature. In this paper, we propose to use Cross-Entropy Monte Carlo (CEME) to find the tau-path with optimal tau-score, the weighted sum of Kendall's tau of reordered accumulating subsets. Specifically, by formulating the problem of finding the optimal tau-path as maximizing tau-score, we explore the usage of a CEMC method for solving such a combinatorial problem. Instead of placing a discrete uniform distribution on all the potential solutions, an iterative importance sampling technique is utilized to slowly tighten the net to place most distributional mass on the optimal solution and its neighbors. Extensive simulation studies were performed to assess the performance of the method. With satisfactory simulation results, the method was applied to the NCI-60 gene expression data to illustrate its utility.

e-mail: *fangyuan.zhang@ttu.edu*

105. STATISTICAL GENETICS: HETEROGENEITY AND HIERARCHY

AGGREGATED QUANTITATIVE MULTIFACTOR DIMENSIONALITY REDUCTION

Rebecca E. Crouch*, University of Kentucky

Katherine L. Thompson, University of Kentucky

Richard J. Charnigo, University of Kentucky

We consider the problem of making predictions for quantitative phenotypes based on gene-to-gene interactions among selected Single Nucleotide Polymorphisms (SNPs). Previously, Quantitative Multifactor Dimensionality Reduction (QMDR) has been applied to detect gene-to-gene interactions associated with high measurements of quantitative phenotypes, by creating a dichotomous predictor (high/low) from one interaction which has been deemed optimal. That method does not take into account cumulative effects from multiple interactions. To address this, we propose an Aggregated Quantitative Multifactor Dimensionality Reduction (AQMDR), which exhaustively considers all k-way interactions among a set of SNPs and replaces the dichotomous predictor from QMDR with a continuous aggregated score. We propose three distinct aggregated scores, which dictate the weight assigned to specific interactions based on p-values from permutation testing. We evaluate this new AQMDR method in a series of simulations for 2-way and 3-way interactions, comparing the new method with the original QMDR, and examining possible advantages of particular proposed aggregated scores. In simulation, AQMDR yields consistently smaller prediction error than QMDR, particularly when more than one significant interaction is present.

e-mail: *rebecca.crouch@uky.edu*

THE INFLUENCE OF POPULATION STRATIFICATION ON GENOMIC HERITABILITY

Gota Morota*, University of Nebraska, Lincoln

The availability of the large volume of molecular markers has

reshaped the landscape of statistical approaches to characterize population structure. Here we sought to apply a reparameterized genomic best linear unbiased prediction (GBLUP) model to infer the impact of population stratification on the estimates of genomic heritability using single-nucleotide polymorphisms. With this GBLUP, a phenotype is regressed on eigenvectors extracted from a genomic relationship matrix, and genomic heritability is expressed as a function of regression coefficients for eigenvectors. The influence of eigenvectors was quantified with hot carcass weight in admixed beef cattle, male flowering time in maize, and milk yield in homogeneous dairy cattle populations. With the removal of the first 5, 10, or 20 eigenvectors, the proportion of reduction in genomic heritability relative to the total genomic heritability was highest in the beef cattle and maize data sets. Underlying population structure may inflate the genomic heritability estimates derived from all eigenvectors. The reductions in genomic heritability estimates are conjectured as the results of the correction of such a stratification. The findings revealed that GBLUP has a potential to advance the understanding of the genotype-to-phenotype map under population stratification by expanding the scope of possible approaches.

e-mail: *morota@unl.edu*

AN ADAPTIVE TESTING APPROACH FOR META-ANALYSIS OF GENE SET ENRICHMENT STUDIES

Wentao Lu*, Southern Methodist University

Xinlei Wang, Southern Methodist University

In the field of gene set enrichment analysis (GSEA), meta-analysis has been used to integrate information from multiple studies to improve the power of detecting essential gene sets involved in cancer or other human complex diseases. However, existing methods, Meta-Analysis for Pathway Enrichment (MAPE, Shen and Tseng 2010), may be subject to information loss because of using gross summary statistics for combining end results from component studies. Therefore, we adapt meta-analysis approaches originally developed for genome-wide association studies, which are based on fixed effect (FE) and random effects (RE) models, to integrate multiple GSEA studies. We further propose a mixed strategy via adaptive testing for choosing RE versus FE models to achieve greater statistical efficiency and flexibility. The three methods tend to have much better performance than the MAPE methods, and can be applied to both discrete and continuous phenotypes. Specifically, the performance of the adaptive testing method seems to be the most stable in general situations.

e-mail: *wlu1026@yahoo.com*

MODELING SECONDARY PHENOTYPES CONDITIONAL ON GENOTYPES IN CASE-CONTROL STUDIES

Naomi C. Brownstein*, Florida State University

Wei Xue, University of North Carolina, Chapel Hill

Jianwen Cai, University of North Carolina, Chapel Hill

Eric Bair, University of North Carolina, Chapel Hill

Traditional case-control genetic association studies examine relationships between case-control status and one or more covariates. Investigators now commonly study additional phenotypes and their association with the original covariates as secondary aims. Assessing these associations is statistically challenging, as participants do not form a random sample from the population of interest. Standard methods may be biased and lack coverage and power. In fact, analysis of one or more secondary phenotypes is a nontrivial problem and has spawned a great deal of research in recent years. In this paper, we seek an unbiased and efficient method to model the relationship between a genotype and a secondary phenotype in case-control studies. We propose two methods to analyze secondary phenotypes and apply them to the Orofacial Pain: Prospective Evaluation and Risk Assessment (OPPERA) baseline case-control study. First, we propose permutation-based testing method which controls type I error and has high power. Second, we propose an inverse probability weighting method for estimation and bootstrapping method for standard error estimation. The proposed method performs as well as competitors when they are applicable and provides promising results for outcomes to which other methods do not apply.

e-mail: *naomi.brownstein@med.fsu.edu*

THE PARAMETRIC T-TEST'S LATENT WEAKNESS

Daniel P. Gaile*, University at Buffalo

Jeffrey C. Miecznikowski, University at Buffalo

Evidence that the observed distributions of t-test values are not well approximated by appropriate t-distributions is plentiful in the array based literature. We demonstrate that latent state conditions can be a contributing factor. This latent state problem is exacerbated by: A) test multiplicity across large numbers of manifest assays, each with a latent structure, and B) increased accuracy of the manifest assays to discriminate underlying latent structures. We also demonstrate that the problem extends beyond an array based context. For example, we provide a motivating 'toy' data-set and explain how the parametric t-test quantifies the evidence against the null hypothesis as approximately 12.5 million to 1 when it should be quantified as approximately 250 to 1. This result is relevant in many modern experimental settings, such as pilot array / next-generation sequencing studies, where an underlying latent structure is either known to

be true (e.g., methylation and array comparative genomic hybridization) or plausible (e.g., down/up-regulated gene networks). Our findings are also applicable to small animal studies (e.g., mouse and rat studies), for which latent state biological mechanisms are often plausible and the parametric t-test is often applied.
e-mail: d*pgaile@buffalo.edu*

NanoStringDiff: A NOVEL STATISTICAL METHOD FOR DIFFERENTIAL EXPRESSION ANALYSIS BASED ON NanoString nCOUNTER DATA

Hong Wang*, University of Kentucky
Craig Horbinski, Northwestern University
Hao Wu, Emory University
Yinxing Liu, University of Kentucky
Shaoyi Sheng, Paul Laurence Dunbar High School
Arnold J. Stromberg, University of Kentucky
Chi Wang, University of Kentucky

It is expected that the uptake of the advanced medium-throughput NanoString nCounter technology will dramatically increase for mRNA or miRNA differential expression studies in the near future due to its potential advantages including direct measurement of molecule expression levels without amplification, digital readout, and lower cost compare to second-generation sequencing. However, the analysis of nCounter data is hampered because most methods developed for the nCounter system are based on t-tests, which do not fit the nature of the data that is in counts. We develop a novel statistical method for differential expression analysis based on data generated from the NanoString nCounter system. The method considers a generalized linear model of the negative binomial family to characterize count data and allows for multi-factor design. Data normalization is incorporated in the model framework by including data normalization parameters estimated from positive controls, negative controls and housekeeping genes embedded in the nCounter system. We propose an empirical Bayes shrinkage approach to estimate the dispersion parameter and a likelihood ratio test to identify differentiallyexpressed genes. Simulations and real data analysis demonstrate that the proposed method performs better in differential expression detection compared to existing methods.
e-mail: *hong.wang@uky.edu*

MEASUREMENT ERROR IN TESTS FOR GENE-ENVIRONMENT INTERACTIONS: IMPLICATIONS OF GENE-ENVIRONMENT DEPENDENCE

Stacey Alexeeff*, Kaiser Permanente Division of Research
Xihong Lin, Harvard School of Public Health

Gene-environment (G-E) interactions play an important role in studying complex disease etiology. However, measurement error is a concern in assessment of environmental exposures. As motivation, we consider a study testing for an interaction effect of body mass index (BMI) and APOE genotype on plasma cholesterol levels in elderly men, where G-E dependence may be present between BMI and APOE. We study the properties of G-E interaction tests where the environment factor is measured with additive error, deriving analytic closed-form solutions for the bias in linear model regression coefficients. Our main finding is that G-E dependence introduces different effects from those in the classical measurement error literature. Specifically, we show that ignoring measurement error results in (1) more complex bias functions where biases may be toward or away from the null, and (2) tests that may not preserve the type I error rate, leading to a higher rate of spurious associations. We identify specific cases of practical interest when type I error rates will be preserved. We also investigate the performance of regression calibration and SIMEX. We illustrate the proposed methods in simulation studies and an analysis of data from the Normative Aging Study.
e-mail: *Stacey.Alexeeff@kp.org*

106. VARIABLE SELECTION

DATA-DRIVEN CONFOUNDER SELECTION VIA MARKOV AND BAYESIAN NETWORKS

Jenny Haggstrom*, Umea University

To unbiasedly estimate a treatment effect on an outcome unconfoundedness is often assumed. If there is sufficient knowledge of the underlying causal structure then existing confounder selection criteria can be used to select subsets of the observed pre-treatment covariates, X, sufficient for unconfoundedness, if such subsets exists. Here, estimation of these target subsets is considered when the underlying causal structure is unknown. The proposed method is to model the causal structure by a probabilistic graphical model, identify the graph from observed data and select the target subsets given the estimated graph. Using Markov and Bayesian networks the approach is evaluated by simulation both in a setting where unconfoundedness holds given X and in settings where unconfoundedness only holds given subsets of X. Several common target subsets are investigated and the selected subsets are compared with respect to accuracy in estimating the average treatment effect. The proposed method is used with existing software that easily can handle high-dimensional data, in terms of large samples and large number of variables. The results from the simulation study suggest that this approach is suitable when

the sample size is relatively large (>1000) and that certain target subsets yield better results than others.
e-mail: *jenny.haggstrom@umu.se*

FEATURE SELECTION FOR COMPLEX METABOLITE NETWORK

Qingpo Cai*, Emory University
Jian Kang, University of Michigan
Tianwei Yu, Emory University

Metabolomics is the comprehensive analysis of metabolites in the biological system and is becoming one of the major areas in high-throughput biology. Liquid chromatography-mass spectrometry (LC-MS) is a major tool in the metabolomics field. One of the main challenges in analyzing LC-MS data is the lack of the feature identity in the data. Given current technology, a feature can be matched to one or a few metabolites with high confidence. At the same time, some known metabolites are missing in the measurements. Current network/pathway analyses usually ignore the uncertainty in the matching and the missing observations, which could result in erroneous statistical inference. To address the aforementioned issues, in this work, we propose a flexible feature selection framework for complex metabolite network. We adopt a sequential feature screening procedure to select important sub-network and identify the optimal matching. The methods are illustrated via extensive simulation studies and analyses of metabolic profiling data from the Emory Predictive Health cohort.

e-mail: *qingpo.cai@emory.edu*

SEQUENTIAL MULTIPLE TESTING FOR VARIABLE SELECTION

Xinping Cui*, University of California, Riverside
Hailu Chen, University of California, Riverside

Lockhart et al. (2014) proposed a simple covariance test for testing the significance of the predictor variable that enters the current lasso model along the lasso solution path. In this paper, we propose a hybrid sequential multiple testing procedure using covariance test p-values, which has a good power properties with error rate controlled at desired level. Specially, we consider the full underlying hypotheses and the error rate control within each step as well as across all steps along the LASSO solution path. To control FWER at desired level, we propose hybrid-Bonferroni, Hochberg and Sime's methods and compare with Single hypothesis. To control FDR, we propose hybrid-BH and compare it with Single hypothesis BH method and StrongStop algorithm. Simulation studies show that our proposed procedures have higher power with both FWER and FDR controlled at desired level. We also apply the proposed procedure for genetic real data analysis to evaluate our method and compare with other methods discussed above.

e-mail: *xinping.cui@ucr.edu*

VARIABLE SELECTION FOR MODEL-BASED CLUSTERING OF FUNCTIONAL DATA

Kyra Singh*, University of Rochester
Tanzy Love, University of Rochester

In studying the health effects of radiation, clustering techniques to identify subpopulations with densely sampled functional data are important for detecting late effects of radiation. However, extraneous variables can mask the true group structure. Utilizing a variable selection technique is particularly important in model-based clustering where there is little or no a priori knowledge of the structure or number of groups within the data and when there is a large number of variables to consider. Little work on variable selection methods for model-based clustering has been applied to functional data. We propose a greedy search algorithm to integrate variable selection into the clustering procedure, as in "Variable Selection for Model-Based Clustering" (Raftery and Dean, 2006) for functional data. At each step, two models are compared using the BIC. One difficulty in implementing this approach is the lack of software available for constructing multivariate fully functional linear models of functional data represented by splines. We avoid this obstacle by creating a full model using a series of univariate partial regressions with the R package 'fda'. Our new method successfully identifies the most important variables for clustering.

e-mail: *kyra_singh@urmc.rochester.edu*

BAYESIAN VARIABLE SELECTION INCORPORATING BIOLOGICAL PATHWAY INFORMATION USING DEPENDENT SHRINKAGE PRIORS

Changgee Chang*, Emory University
Suprateek Kundu, Emory University
Qi Long, Emory University

Recently, substantial effort has been made in the literature in order to incorporate biological pathway information between genes into the variable selection problems. We propose a Bayesian variable selection model that regularizes the regression coefficients via shrinkage priors where the priors are adaptive to the pathway information. The EM algorithm fitting this model and the oracle property will also be presented. We then show the result of our simulation study where our method outperforms other preexisting methods in terms of not only the performance in variable selection and prediction but also the scalability up to more than tens of thousands variables. Finally we describe the use of our method on real data example.

e-mail: *changgee.chang@emory.edu*

BAYESIAN SPATIAL FEATURE SELECTION FOR MASSIVE NEUROIMAGING DATA VIA THRESHOLDED GAUSSIAN PROCESSES

Ran Shi*, Emory University

Jian Kang, University of Michigan

Motivated by the needs of selecting important features for massive neuroimaging data, we propose a spatially varying coefficient model (SVCMs) with sparsity and piecewise smoothness imposed on the coefficient functions. A new class of nonparametric priors is developed based on thresholded Gaussian processes (TGP). We show that the TGP has a large support on a space of sparse and piecewise smooth functions, leading to posterior consistency in coefficient function estimation and feature selection. Also, we develop a method for prior specifications of thresholding parameters in TGPs. Efficient posterior computation algorithms are developed by adopting a kernel convolution approach, where a modified square exponential kernel is chosen taking the advantage that the analytical form of the eigen decomposition is available. Based on simulation studies, we demonstrate that our methods can achieve better performance in estimating the spatially varying coefficient. Also, the proposed model has been applied to an analysis of resting state functional magnetic resonance imaging (Rs-fMRI) data from the Autism Brain Imaging Data Exchange (ABIDE) study.

e-mail: *rshi3@emory.edu*

107. STATISTICAL AND COMPUTATIONAL CHALLENGES IN OMICS DATA INTEGRATION

DATA INTEGRATION USING NETWORK ANALYSIS AND KERNEL MACHINE METHODS

Katerina Kechris*, University of Colorado, Denver

Dominik Reinhold, University of Colorado, Denver

Junxiao Hu, University of Colorado, Denver

Debashis Ghosh, University of Colorado, Denver

To identify molecular pathways and characterize new subtypes for Chronic Obstructive Pulmonary Disease (COPD), different omics data sets have been generated on subjects from the COPDGene genetic epidemiology study. Analyses on each of the individual data types (e.g., transcript expression, protein biomarkers, metabolites) and pairwise interactions (e.g., transcript and metabolites, genotypes and biomarkers) have identified candidate features for further investigation. We hypothesize that a model, which allows for nonlinear effects and more complicated interactions within proposed pathways, will improve the identification of features associated with COPD signatures. We present results of applying weighted correlation network analysis (WGCNA) to construct modular networks using the omics data, and selecting modules and highly connected hub genes as candidate pathways. Then, kernel machine methods are implemented and developed to test for nonlinear associations with the COPD phenotypes by incorporating the module network structure. This framework is extended for multi-dimensional phenotypes and different types of omics data to define predictive pathways. Strategies for incorporating network structure into kernel machine methods are outlined. Our research was developed as part of a working group on Data Integration sponsored by the Statistical and Applied Mathematical Sciences Institute (SAMSI).

e-mail: *katerina.kechris@ucdenver.edu*

DISCOVERY OF NOVEL LOCI ASSOCIATED WITH COPD BY POOLING INFORMATION FROM RELATED CLINICAL FEATURE AND FUNCTIONAL ANNOTATION

Jiehuan Sun*, Yale School of Public Health

Qiongshi Lu, Yale School of Public Health

Russell P. Bowler, National Jewish Health

Katerina J. Kechris, University of Colorado, Denver

Hongyu Zhao, Yale School of Public Health

In the COPD gene project, standard Genome Wide Association Studies (GWAS) using case-control status has been conducted to identify genetic variants associated with COPD, but only a small number of loci have met statistical significance. The COPD patients were followed longitudinally with their clinical information recorded. Most clinical information can be, but has not been, employed to infer loci associated with COPD. Here, we will use the number of exacerbations for each COPD patient over time as a longitudinal trait to find loci that might be associated with COPD and employ functional annotation to boost the power.

e-mail: *jiehuan.sun@yale.edu*

BAYESIAN MULTIVARIATE MODELING OF THE SPHINGOLIPID PATHWAY

Christine B. Peterson*, Stanford University

Elin B. Sellers, Rice University

Francesco C. Stingo, University of Texas MD Anderson Cancer Center

Marina Vannucci, Rice University

We utilize Bayesian multivariate modeling to elucidate the role of gene expression and metabolite abundances in the sphingolipid pathway, an important biological pathway whose dysregulation has been linked to chronic obstructive pulmonary disease (COPD) phenotypes. Specifically, we take a multi-pronged approach to better understand mechanisms contributing to emphysema, applying Bayesian variable selection for probit regression to identify genes relevant to emphysema status and Bayesian graphical modeling to infer differential metabolic and gene expression networks for subjects with varying levels of emphysema severity. In this talk, I will address

both the improvements to computational scaling required to carry out the analysis as well as the biological significance of the findings.
e-mail: *cbpeterson@gmail.com*

LEVERAGING MULTIPLE OMICS DATA TO INFER PATHWAY DISTURBANCE IN COMPLEX DISEASES

Yuping Zhang*, SAMSI TCGA Data Integration Working Group, University of Connecticut

Complex diseases such as cancer are usually the consequences of accumulated disturbance of pathways instead of individual contributions of single regulatory elements. Recent advances in high-throughput biotechnologies have generated diverse types and huge amounts of data, which provide unprecedented opportunities to systematically investigate pathway-based disturbances of biological systems and their potential consequences in disease development and progression. In this talk, we will describe specific scientific contexts in cancer research, give an example using the Cancer Genome Atlas (TCGA) data set, address statistical issues, and introduce the research efforts by the SAMSI TCGA data integration working group.
e-mail: *yuping.zhang@uconn.edu*

INTEGRATING CLINICAL AND MOLECULAR DATA FOR SURVIVAL PREDICTION IN TCGA

Bin Zhu*, National Cancer Institute, National Institutes of Health

Nan Song, NSABP Foundation

Ronglai Shen, Memorial Sloan Kettering Cancer Center

Veera Baladandayuthapani, University of Texas MD Anderson Cancer Center

Katerina Kechris, University of Colorado, Denver

Hongyu Zhao, Yale University

Large-scale comprehensive molecular profiling of tumor samples has revealed substantial inter-patient heterogeneity in their genomic background and tumor characteristics with unprecedented details. However, very few studies have systematically studied the effect of integrating multiple platforms for patient survival prediction, and the performance of molecular predictors in relation to clinico-pathological variables such as age, stage, and tumor size. We present a kernel-fusion Cox regression framework for integrating clinical and molecular data at the genomic, epigenomic and transcriptomic level to achieve enhanced precision in survival prediction. The kernel fusion approach allows powerful non-linear discrimination based on multiple molecular data types (somatic mutation, DNA copy number, DNA methylation and mRNA expression) and multiple views (e.g., gene, isoform, exon expression). In this talk, we present examples using the TCGA multi-platform molecular data sets. We demonstrate that a) integrating multiple molecular platforms leads to improved prediction accuracies than using any single platform alone; b) for complex tumor types such as lung adenocarcinoma, kernel methods effectively aggregate numerous small effects toward better prediction. Taken together, we present a framework that provide a multi-platform learning of TCGA data sets to significantly improve the survival prediction.
e-mail: *bin.zhu@nih.gov*

108. STATISTICAL METHODS TO HANDLE TREATMENT CROSSOVER AND SWITCH TO AN ALTERNATIVE THERAPY IN RANDOMIZED CONTROLLED TRIALS

REGRESSION BASED IMPUTATION ANALYSIS ADJUSTING FOR SUBSEQUENT THERAPY

Chengqing Wu*, Celgene

Xiaolong Luo, Celgene

Mingyu Li, Celgene

Qiang Xu, Celgene

Guang Chen, Celgene

Bruce E. Dornseif, Celgene

Markus F. Renschler, Celgene

Gary Koch, University of North Carolina, Chapel Hill

Subsequent therapy in clinical trials may confound the analysis of a long term endpoint.We proposed a regression based imputation procedure to mediate the confounding effect and to allow inferences about the treatment e.ect due to the originally assigned treatments. The procedure is evaluated by intensive simulation analyses. The proposed method has been applied to an analysis of a multicenter, randomized, open-label, phase 3 trial conducted to evaluate azacitidine efficacy and safety vs conventional care regimens (CCRs) in elderly patients with newly diagnosed acute myeloid leukemia (AML) with >30% bone marrow blasts.
e-mail: *cwu@celgene.com*

WEIGHTED LOGRANK TESTS FOR TREATMENT EFFECTS IN CLINICAL TRIALS WITH CROSSOVER

Rajeev Ayyagari*, Analysis Group, Inc.

James M. Robins, Harvard School of Public Health

In a recent randomized trial comparing the efficacy and safety of an active study drug to placebo in patients with advanced renal cell carcinoma, several patients in the placebo arm received the drug

upon progression. Many patients in the active arm discontinued active therapy after progression. Due to this two-way crossover, the observed effect of the drug on overall survival was attenuated, and the logrank test of differences between the two arms was not significant. To address this loss of power, a family of weighted logrank tests was developed using time-dependent weights proportional to the difference between the two arms in actual proportion of patients receiving pazopanib. We report the results of a simulation study evaluating the type I error and power of the weighted logrank test under varying assumptions on confounding and the proportion of patients who cross over in each arm.

e-mail: *rajeev.ayyagari@analysisgroup.com*

ON THE USE OF RANK-PRESERVING STRUCTURAL FAILURE TIME MODEL TO ACCOUNT FOR BOTH TREATMENT CROSS-OVER AND SWITCH TO ALTERNATIVE THERAPIES

Liewen Jiang*, Biostatistics, Infinity Pharmaceuticals, Inc.

Shijie Tang, Biostatistics, Infinity Pharmaceuticals, Inc.

Lingling Li*, Harvard University

The work is motivated by the analytic challenges encountered in a Phase 3 randomized clinical trial to compare the efficacy between a study drug and an active control on overall survival among patients with relapsed or refractory Chronic Lymphocytic Leukemia (CLL) or Small Lymphocytic Lymphoma (SLL). Patients are allowed to crossover to the other treatment arm after disease progression. In addition, patients may switch to an alternative therapy which has been demonstrated to have better efficacy than the active control during the course of the study. In theory, the RPSFT model can be used to handle both treatment crossover and switch by specifying multiple parameters in the structural model. However, in practice, it is very difficult to estimate more than one parameter using randomization alone. We propose a strategy to reduce the dimensionality of the parameter space by imposing additional parametric assumptions on the likelihood and then estimating the nuisance parameters via fitting the observed data to the specified likelihood. We obtain the final estimate of the parameter of interest by plugging in the nuisance parameter estimates to the estimating equations from the original RPSFT model. We will present the results from a comprehensive simulation study to evaluate the performance of the extended RPSFT approach. We design the simulation study based on the empirical setting in the motivational Phase 3 trial. In addition, we will compare its performance to the IPCW approach which imposes fewer assumptions and thus may be more robust but may also be less efficient.

e-mail: *Liewen.Jiang@infi.com*

BIOLOGY, CAUSAL MODELS, AND CROSS-OVER IN CANCER TRIALS

James M. Robins*, Harvard School of Public Health

I will review various causal models and associated instrumental variables estimation methods when, after disease progression, subjects in a cancer trial can cross over to the treatment assigned in the opposite arm. The focus will be on the biological plausibility of the models. Biological plausibility is essential because, with instrumental variables, identification of causal effects is through the functional form of the causal model

e-mail: *robinsjami1@gmail.com*

109. SENTINEL STATISTICAL METHODS WORKING GROUPS, CHALLENGES WITH USING CLAIMS DATA FOR PUBLIC HEALTH

INTRODUCTION TO SENTINEL DISTRIBUTED DATA SYSTEM AND SELECTED METHODS WORK

Judith C. Maro*, Harvard Medical School and Harvard Pilgrim Health Care Institute

In this talk, we first introduce the Sentinel distributed data system and common data models that allow data partners to keep full control of their own data while contributing useful information to important public health and research studies. This system has been adopted by other large, national initiatives such as the PCORnet and NIH Collaboratory Distributed Research Network. We then give an overview of completed, ongoing, and planned statistical methods work in Mini-Sentinel (the pilot phase of Sentinel) and Sentinel. We have developed and implemented a variety of methods for all three steps of active surveillance: signal generation, signal refinement, and signal evaluation. In the end, we will discuss the current statistical challenges and the planned methods and tool development to further enhance the functionality of the Sentinel system.

e-mail: *jmaro@mit.edu*

LESSONS LEARNED FROM TWO SENTINEL SEQUENTIAL SURVEILLANCE ACTIVITIES: SAXAGLIPTIN AND RIVAROXABAN

Bruce Fireman*, Kaiser Permanente Division of Research

Sentinel conducted surveillance of myocardial infarction in saxagliptin users and surveillance of ischemic and bleeding outcomes in rivaroxaban users. Both of these surveillance activities tested multiple hypotheses at multiple interim analyses. The statistical threshold for a "signal" was adjusted for the sequential testing of each hypothesis but not for the multiplicity of hypotheses. In both surveillance activities the threshold for a signal was crossed at an

interim analysis by a Wald test statistic yielded by a stratified Cox regression comparing users of the target drug with propensity-score matched users of an active comparator. After the "signals", later analyses re-examined the accumulating evidence and assessed possible sources of bias. The challenges included late-arriving data and revised data that left later analyses somewhat less anchored to earlier analyses than anticipated, and privacy-preserving limitations on what individual-level data could be pooled. Here we summarize some of the lessons learned.

e-mail: *bruce.fireman@kp.org*

SURVIVAL METHODS FOR POSTMARKETING MEDICAL PRODUCT SURVEILLANCE IN A DISTRIBUTED NETWORK

Andrea J. Cook*, Group Health Research Institute

Robert Wellman, Group Health Research Institute

Rima Izem, U.S. Food and Drug Administration

Azadeh Shoaibi, U.S. Food and Drug Administration

Ram Tiwari, U.S. Food and Drug Administration

Susan Heckbert, University of Washington

Lingling Li, Harvard University

Rongmei Zhang, U.S. Food and Drug Administration

Jennifer Nelson, Group Health Research Institute

Conducting observational postmarketing medical product safety surveillance is important for detecting rare adverse events not identified pre-licensure. New systems for the safety surveillance setting have been built using electronic healthcare data that keeps the individual patient data within the health plan and establishes a distributed data network to share deidentified data to answer important safety questions about new medical products. One such network is the FDA Sentinel Initiative. We will present survival methods tailored to these networks to estimate hazard ratios using different Cox Proportional Hazard models that account for confounding using approaches such as regression, stratification and exposure matching. To assess the performance of such methods we will conduct a simulation study comparing methods in terms of bias, power, and coverage. We will focus our simulation comparison on the rare event setting with strong across health plan/site confounding due to differential uptake of new medical products.

e-mail: *cook.aj@ghc.org*

110. STATISTICAL MODELING OF DATA ON HEALTH POLICY AND COST

ON STATISTICAL MODELING OF NATIONAL SURVEYS TO ASSESS THE IMPACT OF STATE SPECIFIC MEDICAL MARIJUANA POLICIES

Christine Mauro, Columbia University

Melanie M. Wall*, Columbia University and New York State Psychiatric Institute

Since 1996, 23 states have passed laws legalizing medical use of marijuana, and other states are considering such laws. Using aggregated data from two national surveys, we showed that states with medical marijuana laws (MML) had significantly higher prevalence of adolescent (Wall et al., 2011) and adult marijuana use (Cerda et al., 2012) than other states. But due to limitations in the data sources, it was not possible to look at changes within states from pre to post law passage. Using individual level data from two different repeated cross-sectional national surveys, Monitoring the Future and the National Survey on Drug Use and Health we will investigate whether passage of state medical marijuana laws leads to a change in illicit marijuana use within the states that pass them. The present talk will elaborate the multilevel statistical modeling developed to address the question of within state changes in illicit marijuana use. Specifically we will describe issues related to: smoothly controlling for secular changes, dealing with sampling weights that were not originally created for state level analyses, incorporating random effects for individual covariates at the state level, and graphically presenting results that meaningfully capture state-to-state trends and differences.

e-mail: *mmwall@columbia.edu*

AN IMPROVED SURVIVAL ESTIMATOR FOR MEDICAL COSTS WITH CENSORED DATA USING KERNEL METHODS

Shuai Chen*, University of Wisconsin, Madison

Wenbin Lu, North Carolina State University

Hongwei Zhao, Texas A&M Health Science Center

Costs assessment and cost-effectiveness analysis serve as an essential part in economic evaluation of medical interventions. In clinical trials and many observational studies, costs as well as survival data are frequently censored. Standard techniques for survival-type data are often invalid in analyzing censored cost data, due to the induced dependent censoring problem. In this talk, we will first examine the equivalency between a redistribute-to-the right (RR) algorithm and the popular Kaplan-Meier method for estimating the survival function of time. Next, we will extend the RR algorithm to the problem of estimating the survival function of medical costs,

and discuss RR-based estimators. Finally, we will propose a kernel-based estimator for the survival function of costs, which is shown to be monotone, consistent, and more efficient than some existing survival estimators. We conduct simulation experiments to compare these survival estimators for costs and apply them to a data example from a randomized cardiovascular clinical trial.

e-mail: *schen264@wisc.edu*

COMPARISON IN MEDICAL COST BETWEEN A CANCER SURVIVOR COHORT AND THE GENERAL POPULATION USING LONGITUDINAL PHYSICIAN CLAIMS

Huijing Wang*, Simon Fraser University

X. Joan Hu, Simon Fraser University

The population of cancer survivors has been increasing rapidly as a result of advances in treatment. They are often at risk of subsequent and ongoing health problems that are primarily treatment-related. Risk assessment of the later effects and comparison to the general population are of much interest to policymakers, care providers, and the survivors and their families. This talk compares the longitudinal medical costs associated with physician claims over time between a cohort of cancer survivors diagnosed before the age of 20 and a random selected general population sample with matching sex and birth-year. Firstly, we conduct the conventional longitudinal analyses on the survivor cohort and general population separately or together by GEE. The difference over time between the cohort and the population is found via time-varying effects. Secondly, we employ a latent class model to formulate the longitudinal medical costs of the cohort into two latent classes, at-risk and not-at-risk groups, and conduct the associated inference assuming the not-at-risk group has the same distribution of medical costs as the general population. Further, the latent class model allows identifying risk factors of the survivors to subsequent and ongoing problems. The approach can thus provide risk classification and prediction for the cancer survivors.

e-mail: *hwa40@sfu.ca*

"NONPARAMETRIC" META ANALYSIS WITH UNKNOWN STUDY-SPECIFIC PARAMETERS AND WITH AN APPLICATION TO HEALTH POLICY DATA

Min-ge Xie*, Rutgers University

Meta-analysis is a valuable tool for combining information from independent studies in health care studies and other fields. However, most common meta-analysis techniques rely on distributional assumptions that are difficult, if not impossible, to verify. For instance, in the commonly used fixed-effects and random-effects models, we take for granted that the underlying study-level parameters are either exactly the same across individual studies or that they are realizations of a random sample from a population, often under a parametric distributional assumption. In this talk, we present a new framework for summarizing information obtained from multiple studies and make inference that is not dependent on any distributional assumption for the study-level parameters. Specifically, we assume the study-level parameters are unknown, fixed parameters and draw inferences about, for example, the quantiles of this set of parameters using study-specific summary statistics. This type of problem is known to be quite challenging in statistical inference (c.f., Hall & Miller, 2010). We utilize a novel resampling method via the confidence distributions of the study-level parameters to construct confidence intervals for the above quantiles. We justify the validity of the inference procedure asymptotically and compare the new procedure with the standard bootstrapping method. We also illustrate our proposal with simulations and real data related to health care policy studies (Joint work with Brian Claggett and Lu Tian).

e-mail: *mxie@stat.rutgers.edu*

111. WEIGHT MODIFICATION IN SAMPLE SURVEYS

WEIGHT MODIFICATION IN SAMPLE SURVEYS: AN OVERVIEW

Malay Ghosh*, University of Florida

This talk overviews weighting methods in sample surveys. The setting is the analysis of data from surveys involving complex probability sampling designs, potentially with nonresponse and auxiliary information. Interest concerns inference about finite population quantities, such as population totals and means, or by extension, functions of totals such as ratios. I will discuss advantages as well as limitations of some of the standard estimators such as those proposed by Horvitz-Thompson and Hajek.

e-mail: *ghoshm@stat.ufl.edu*

WEIGHT TRIMMING AND WEIGHT SMOOTHING PROCEDURES FOR SURVEY DATA

David Haziza*, Université de Montréal

It is well known that point estimators tend to be unstable when the survey weights are highly dispersed and exhibit a low correlation with the study variables. This problem was nicely illustrated by Basu (1971) with his famous example of circus elephants. To limit the impact of highly dispersed weights, a number of techniques has been proposed in the literature, including weight trimming and weight smoothing methods. Although both types of methods are different in nature, they share the same goal: modify the survey weights so that the resulting estimators have a lower mean square error than that of the usual estimators (e.g., the Horvitz-Thompson estimator). Reduction of the mean square error is

generally achieved at the expense of introducing a bias. Therefore, the treatment of survey weights by either weight trimming or weight smoothing methods can be viewed as a compromise between bias and variance. In this talk, we will review several weight trimming methods and discuss their properties.

e-mail: *david.haziza@umontreal.ca*

WEIGHT MODIFICATION IN SAMPLE SURVEYS: USING REGRESSION MODELS

Qixuan Chen*, Columbia University

For inferences from samples selected from finite populations, weights are typically used to mitigate the effects of selection bias and nonresponse and to improve the quality of the inferences using auxiliary information. Although standard weighted estimators have many desirable properties, cases in which weights are both highly variable and weakly correlated with the study variables can yield estimators with poor mean square errors. In this talk, I will discuss two approaches to modifying survey weights based on regression models. I will first review methods that modify weights arising from models for the survey variable, with the survey weights treated as covariates. Such methods include "weight pooling" methods (Elliott and Little, 2000; Elliott, 2008; 2009), "weight smoothing" methods (Elliott and Little, 2000; Elliott, 2007), and the penalized spline predictive methods (Zheng and Little, 2003; 2005; Chen, Elliott and Little, 2010; 2012). I will then review weight modeling methods that replace the survey weights by the predictions obtained by modeling the survey weight conditionally on the survey variables (Beaumont, 2008; Kim and Skinner, 2013). I will discuss the properties of the methods in both approaches.

e-mail: *qc2138@cumc.columbia.edu*

112. GENERALIZING CLINICAL DATA ACROSS STUDIES/POPULATIONS

ADJUSTED COMPARISONS TO EXTERNAL CONTROLS USING BOTH INDIVIDUAL PATIENT DATA AND PUBLISHED SUMMARY STATISTICS

James E. Signorovitch*, Analysis Group Inc.

David Cheng, Harvard University

Comparisons of treatment outcomes to an external control group are often needed to assess comparative efficacy and safety. This is common, for example, when uncontrolled trials are conducted for breakthrough therapies, and in rare diseases and late stage oncology settings where practical and ethical considerations preclude randomized internal control groups. When comparing to external controls, adjustment for baseline difference between treatment groups is of paramount importance. Traditional approaches to adjustment, such as matching, require individual patient data from all treatment groups. However, individual patient data are not always accessible for all external control groups of interest. In particular, placebo or active therapy arms from recent clinical trials are often of high interest as external controls, but the individual patient data are typically proprietary. We describe an extension of propensity score weighting that can be used to adjust for multiple baseline characteristics when only published aggregate data are available for external controls. Advantages and limitations compared to regression-based approaches will be described, along with example applications to single-arm trial data. New results on bias-corrected variance estimation in small sample settings, which are common for single-arm studies, will be described through a simulation study.

e-mail: *james.signorovitch@gmail.com*

ROBUST METHODS FOR TREATMENT EFFECT CALIBRATION, WITH APPLICATION TO NON-INFERIORITY TRIALS

Zhiwei Zhang*, U.S. Food and Drug Administration

Lei Nie, U.S. Food and Drug Administration

Guoxing Soon, U.S. Food and Drug Administration

Zonghui Hu, National Institute of Allergy and Infectious Diseases, National Institutes of Health

In comparative effectiveness research, it is often of interest to calibrate treatment effect estimates from a clinical trial to a target population that differs from the study population. One important application is an indirect comparison of a new treatment with placebo on the basis of two separate clinical trials: a non-inferiority trial comparing the new treatment with an active control and a historical trial comparing the active control with placebo. The available methods for treatment effect calibration include an outcome regression (OR) method based on a regression model for the outcome and a weighting method based on a propensity score (PS) model. This article proposes new methods for treatment effect calibration: one based on a conditional effect (CE) model and two doubly robust (DR) methods. The first DR method involves a PS model and an OR model, is asymptotically valid if either model is correct, and attains the semiparametric information bound if both models are correct. The second DR method involves a PS model, a CE model and possibly an OR model, is asymptotically valid under the union of the PS and CE models, and attains the semiparametric information bound if all three models are correct. The various methods are compared in a simulation study and applied to recent clinical trials for treating human immunodeficiency virus infection.

e-mail: *zhiwei.zhang@fda.hhs.gov*

SENSITIVITY ANALYSIS FOR AN UNOBSERVED MODERATOR IN RCT-TO-TARGET POPULATION GENERALIZATION OF TREATMENT EFFECT

Trang Q. Nguyen*, Johns Hopkins Bloomberg School of Public Health

Cyrus Ebnesajjad, Johns Hopkins Bloomberg School of Public Health

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

Elizabeth A. Stuart, Johns Hopkins Bloomberg School of Public Health

Due to treatment effect heterogeneity, a treatment's effect in a randomized controlled trial (RCT) may differ from its effect if applied to a target population of interest. An estimate of the latter can be obtained "if all effect moderators are observed in the RCT and in a dataset representing the target population" by adjusting for the difference in the moderators' distribution between the two samples. We consider sensitivity analyses for two situations: (1) where we cannot adjust for a moderator V observed in the RCT because it is not observed in the target population; and (2) where we worry that treatment effect may be moderated by factors not observed even in the RCT, which we represent as a composite moderator U. Assuming an additive potential outcomes model, for situation (1), we offer: (i) a bias-formula-based technique that involves specifying a range for the target population mean of V; (ii) a weighting-based technique that involves specifying a range for the target population distribution of V given observed moderators Z (if any); and (iii) a hybrid technique combining weighting and bias formula elements. For situation (2), we offer modified versions of options (i) and (iii) for a U uncorrelated with Z and possibly other covariates.

e-mail: *tnguye28@jhu.edu*

BAYESIAN NETWORK META-ANALYSES OF MULTIPLE DIAGNOSTIC TESTS

Haitao Chu*, University of Minnesota

Xiaoye Ma, Amgen Inc.

Qinshu Lian, University of Minnesota

Yong Chen, University of Pennsylvania

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

In studies evaluating the accuracy of diagnostic tests, three designs are commonly used: (1) the crossover design; (2) the randomized design; and (3) the non-comparative design. Existing methods on meta-analysis of diagnostic tests mainly considered the simple case when the reference test in all or none of studies can be considered as a gold standard test, and when all studies use either a randomized or non-comparative design. Yet the proliferation of diagnostic instruments and diversity of study designs being used have boosted the demand to develop more general methods to combine studies with or without a gold standard test using different designs. In this talk, we discuss two related frameworks from the missing data perspective for network meta-analysis of diagnostic tests to compare multiple tests simultaneously. It accounts for the potential correlation between multiple tests within a study and heterogeneity across studies. Our model is evaluated through simulations and illustrated using a real data analysis.

e-mail: *chux0051@umn.edu*

113. NOVEL STATISTICAL METHODS FOR SEQUENCING DATA - FROM QUALITY CONTROL TO FALSE POSITIVES

A GENERALIZED SIMILARITY U TEST FOR MULTIVARIATE ANALYSIS OF SEQUENCING DATA

Changshuai Wei, University of North Texas Health Science Center

Qing Lu*, Michigan State University

Sequencing-based studies are emerging as a major tool for genetic association studies of complex diseases. These studies pose great challenges to the traditional statistical methods because of the high-dimensionality of data and the low frequency of genetic variants. Moreover, there is a great interest in biology and epidemiology to identify genetic risk factors contributed to multiple disease phenotypes. The multiple phenotypes can often follow different distributions, which brings an additional challenge to the current statistical framework. In this paper, we propose a generalized similarity U test, referred to as GSU. GSU is a similarity-based test that can handle high-dimensional genotypes and phenotypes. We studied the theoretical properties of GSU, and provided the efficient p-value calculation for association test as well as the sample size and power calculation for the study design. Through simulation, we found that GSU had advantages over existing methods in terms of power and robustness to phenotype distributions. Finally, we used GSU to perform a multivariate analysis of sequencing data in the Dallas Heart Study and identified a joint association of 4 genes with 5 metabolic related phenotypes.

e-mail: *qlu@epi.msu.edu*

STATISTICAL AND COMPUTATIONAL ASPECTS IN THE ANALYSIS OF GENOMIC DATA FROM FAMILY BASED DESIGNS

Ingo Ruczinski*, Johns Hopkins Bloomberg School of Public Health

"Family based study designs are regaining popularity because large-scale sequencing can help to interrogate the relationship between disease and variants too rare in the population to be

detected through any test of association in a conventional case-control study, but may nonetheless co-segregate with disease within families. In addition, family based designs can also allow for the assessment of de novo events and parent-of-origin effects. In this presentation, we mainly focus on statistical and computational aspects in the analysis of sequencing and array data from nuclear families with one or more affected probands, with an emphasis on improvements in scalability and variant detection.

e-mail: *ingo@jhu.edu*

EMPIRICAL ESTIMATION OF SEQUENCING ERROR RATES USING SMOOTHING SPLINES

Xuan Zhu, University of Texas MD Anderson Cancer Center

Jian Wang, University of Texas MD Anderson Cancer Center

Bo Peng, University of Texas MD Anderson Cancer Center

Sanjay Shete*, University of Texas MD Anderson Cancer Center

Next-generation sequencing has been used by investigators to address a diverse range of biological problems. However, compared to conventional sequencing, the error rates for next-generation sequencing are often higher, which impacts the downstream genomic analysis. Recently, Wang et al. (2012) proposed a shadow regression approach to estimate the error rates for next-generation sequencing data based on the assumption of a linear relationship between the number of reads sequenced and the number of reads containing errors (denoted as shadows). However, this linear read-shadow relationship may not be appropriate for all types of sequence data. We proposed an empirical error rate estimation approach that employs cubic and robust smoothing splines to model the relationship between the number of reads sequenced and the number of shadows. The simulation results show that the proposed approach provided more accurate error rate estimations than the shadow linear regression approach for all the scenarios tested. We also applied the proposed approach to assess the error rates for the sequence data from several studies.

e-mail: *sshete@mdanderson.org*

RARE VARIANTS ASSOCIATION ANALYSIS IN LARGE-SCALE SEQUENCING STUDIES AT THE SINGLE LOCUS LEVEL

X. Jessie Jeng, North Carolina State University

Z. John Daye, University of Arizona

Wenbin Lu, North Carolina State University

Jung-Ying Tzeng*, North Carolina State University

Genetic association analyses of rare variants in next-generation sequencing (NGS) studies are fundamentally challenging due to the presence of a very large number of candidate variants at extremely low minor allele frequencies. Recent developments often focus on pooling multiple variants to provide association analysis at the gene instead of the locus level. Nonetheless, pinpointing individual variants is a critical goal for genomic researches as such information can facilitate the precise delineation of molecular mechanisms and functions of genetic factors on diseases. Due to the extreme rarity of mutations and high-dimensionality, significances of causal variants cannot easily stand out from those of noncausal ones. Consequently, standard false positive control procedures, such as the Bonferroni and false discovery rate (FDR), are often impractical to apply, as a majority of the causal variants can only be identified along with a few but unknown number of noncausal variants. To provide informative analysis of individual variants in large-scale sequencing studies, we propose the FAlse Negative control Screening (FANS) procedure that can include a large proportion of causal variants with high confidence by introducing a novel statistical inquiry to determine those variants that can be confidently dispatched as noncausal. The FANS provides a general framework that can accommodate for a variety of models and significance tests. The procedure is computationally efficient and can adapt to the underlying proportion of causal variants and quality of significance rankings. Extensive simulation studies across a plethora of scenarios demonstrate that the FANS is advantageous for identifying individual rare variants, whereas the Bonferroni and FDR are exceedingly over-conservative for rare variants association studies. In the analyses of the CoLaus dataset, FANS has identified individual variants most responsible for gene-level significances. Moreover, single-variant results using the FANS have been successfully applied to infer related genes with annotation information.

e-mail: *jytzeng@stat.ncsu.edu*

114. BAYESIAN CAUSAL INFERENCE

UTILIZING VALIDATION DATA: A BAYESIAN VARIABLE SELECTION APPROACH TO ADJUST FOR CONFOUNDING

Joseph Antonelli*, Harvard School of Public Health

Francesca Dominici, Harvard School of Public Health

Large administrative databases are becoming increasingly popular as they allow us to examine a vast number of scientific questions in comparative effectiveness research. Frequently these databases have very large sample sizes, but do not measure many potentially important confounders that are required for valid estimation of a causal effect. However, in many settings these potential confounders are measured in a small sample of the study population from a validation data set. We propose a new Bayesian data augmentation and variable selection approach for estimating an average causal effect in the main study that uses validation data to adjust for

confounding. Our approach identifies the key confounders among all potential covariates, imputes these confounders in the larger study population, and incorporates all of the uncertainty in these two procedures into the final causal effect estimates. We apply our method to the analysis of 26,559 Medicare enrollees hospitalized with a brain tumor to estimate the effect of surgical resection on survival. We use validation data on 4,428 brain tumor patients from SEER, which measured 9 additional potential confounders, and found that the average causal effect is smaller when adjusting for the larger set of covariates.

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

A CAUSAL INFERENCE APPROACH FOR ESTIMATING AN EXPOSURE RESPONSE CURVE: ESTIMATING HEALTH EFFECTS AT LOW POLLUTION LEVELS

Georgia Papadogeorgou*, Harvard University

Francesca Dominici, Harvard School of Public Health

Many methods have been developed to estimate a potentially non-linear exposure-response (ER) curve, while accounting for known observed confounders. However, none of these approaches account for the possibility of different confounding variables at different exposure levels, or for the uncertainty of model selection. Furthermore, it is often the case that the sample size at extreme exposure levels is significantly smaller than at average exposure. Extrapolation and estimation of the ER curve at extreme exposure levels using information from normal levels can lead to significant bias in the estimation of causal effects. Such a situation is met in the study of the health effects of low ambient air pollution. While a lot of information exists for areas of average air pollution, we would like to estimate the causal effect of ambient air pollution at low levels, while using the information of all exposure levels to gain power. Our approach borrows information across exposure levels to identify the important confounding variables at each level separately. Using this information, we estimate the whole ER curve, which will have a causal interpretation, while accounting for the uncertainty in confounder selection at each level of exposure.

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

A SEMI-PARAMETRIC DOUBLE ROBUST BAYESIAN'S APPROACH TO CASUAL INFERENCE

Bin Huang*, Cincinnati Children's Hospital Medical Center

Chen Chen, Cincinnati Children's Hospital Medical Center

Current literature saw heightened effort of incorporating propensity score into Bayesian's approach to causal inference. These work highlighted difficulties of framing Bayesian's double robust approaches similarly as seen in Frequentist approach. Directly joint modeling of propensity score and outcome modeling could produce biased causal estimates due to the "feedback" issue (McCandles 2009, 2010; Zigler et al 2013, 2015). Others (Graham et al 2015) took an approximate Bayesian approach. Whether Saarela et al's (20015) formalization of Bayesian's interpretation to inverse probability weight is full Bayesian is still under active debating. A full Bayesian's approach embedding propensity scores within broader Bayesian modeling strategies, such as BMA and BART, remain to be found. In this study, we propose a semiparametric Bayesian modeling approach that is designed to enjoy double robust property, and offer a way of directly balancing propensity score. Utilizing the simulation design provided in Kang and Schafer (2007), we compared this approach with the existing Bayesian's causal inference methods. The results demonstrate superior performances of the proposed method. We further applied the method to evaluate clinical effectiveness of early aggressive treatment in treating newly onset juvenile arthritis patient population using a registry data.

e-mail: *bin.huang@cchmc.org*

ADDRESSING UNMEASURED CONFOUNDING USING EXTERNAL VALIDATION DATA: IMPROVING BayesPS APPROACH

Negar Jaberasari*, Cincinnati Children's Hospital Medical Center

Bin Huang, Cincinnati Children's Hospital Medical Center

Bayesian's approach to causal inference have the advantage of combining data from different sources. Utilizing external validation data, McCandless et al (2012) proposed to impute a BayesPS score from the propensity score model, to address unmeasured confounders. Their BayesPS is a scalar summary of unmeasured confounders from the propensity score modeling conditioning on the vector of observed covariates. Improving upon their method, we propose a slightly different way of imputing BayesPS score, and introducing an additional propensity score. Investigating the setting where the outcome is continuous and the treatment assignment is binary, our simulation studies demonstrate improved performances in estimating averaged causal effect and the averaged potential outcomes. We also apply the method to a comparative clinical effectiveness study of treating children with newly onset of juvenile idiopathic arthritis.

e-mail: *negar.jaberansari@cchmc.org*

BAYESIAN METHODS FOR MULTIPLE MEDIATORS: PRINCIPAL STRATIFICATION AND CAUSAL MEDIATION ANALYSIS OF POWER PLANT EMISSION CONTROLS

Chanmin Kim*, Harvard University

Michael Daniels, University of Texas, Austin

Joseph Hogan, Brown University

Christine Choirat, Harvard University
Corwin Zigler, Harvard University

One goal of air quality regulatory policies in the US is to reduce emissions that are precursors to the formation of PM2.5 in the atmosphere, which is known to be associated with adverse health outcomes. However, the presumed relationships between scrubbers, emissions, and ambient PM2.5 have never been estimated or empirically verified amid the realities of actual regulatory implementation. The goal of this paper is to develop new statistical methods to quantify these causal relationships. We frame this problem as one of mediation analysis to evaluate the extent to which the causal effect of a scrubber on ambient PM2.5 is mediated through causal effects on power plant emissions. Since power plants emit various pollutants including sulfur dioxide, nitrous oxides and carbon dioxide, we develop new statistical methods for settings with multiple intermediate mediating factors that are measured contemporaneously, may interact with one another, and may exhibit joint mediating effects. Specifically, we propose new methods leveraging two related frameworks for causal inference in the presence of mediating variables: principal stratification and causal mediation analysis. Both approaches are anchored to the exact same models for the observed data, which we specify with flexible Bayesian nonparametric techniques and the common set of identifying assumptions. The principal stratification and causal mediation analyses are interpreted in tandem to provide the first comprehensive empirical investigation of the presumed causal pathways.

e-mail: *ckim@hsph.harvard.edu*

115. BIOMARKERS

ESTIMATING THE RECEIVER OPERATING CHARACTERISTIC CURVE FOR PAIRED FAMILY DATA IN A CASE-CONTROL DESIGN

Yalda Zarnegarnia*, University of Miami
Shari Messinger Cayetano, University of Miami

Receiver operating characteristic (ROC) curve have been widely used in medical research to evaluate the accuracy of diagnostic tests in differentiating between diseased and undiseased groups. However many areas of research involve the analysis of correlated biomarker data where subjects may have same genetic or environmental factors such as familial data. In order to provide the ROC curve, the approaches appropriate to a family matched case-control design, must be able to accommodate the inherent correlation in correctly estimating the biomarker's ability to differentiate between groups. The information about the correlation among the subjects can be helpful to identify family members at increased risk of disease development. This talk will review available methods for ROC curve construction in settings with correlated data and will discuss the limitations of current methods for analyzing correlated paired data from a familial, case-control design. We will present an alternative approach using Conditional ROC curves, to provide appropriate ROC curves for correlated paired data. The proposed approach will take into account correlation between biomarker values, producing ROC curves that specifically illustrate the ability of a biomarker to discriminate between groups in a familial or paired design.

e-mail: *yzarnegarnia@med.miami.edu*

EVALUATING LONGITUDINAL BIOMARKERS

Rosa Oliveira, Instituto Politecnico do Porto, Portugal
Raymond Carroll, Texas A&M University
Armando Teixeira-Pinto*, University of Sydney

Many studies on the diagnostic or prognostic ability of clinical biomarkers focuses on the evaluation of a single measurement of a physiological parameter and its association with an outcome of interest. However, there are settings where the longitudinal profile of a biomarker may be more informative than the single value of a cross-sectional observation. Motivated by an example evaluating C-reactive protein (CRP) as a marker of Sepsis resolution in patients admitted to intensive case, we discuss how the CRP trend, over several days, can be used to model the probability of death for these patients. We propose a two-stage approach that first models CRP as a linear trajectory and then uses the estimated intercepts and slopes as predictor variables in the mortality model. Given that these two quantities are estimates of the true intercepts and slopes, we show that their direct use in the mortality model produces biased estimates of their effect. By recognizing this setting as a measurement error problem, we studied a pseudo-likelihood approach and adapted the regression calibration method to reduce the bias in the two-step approach. Results from a monte carlo simulation as well as results from a real data are discussed.

e-mail: *armando.teixeira-pinto@sydney.edu.au*

EVALUATION OF BIOMARKERS FOR TREATMENT SELECTION USING INDIVIDUAL PARTICIPANT DATA META-ANALYSIS

Chaeryon Kang*, University of Pittsburgh
Holly Janes, Fred Hutchinson Cancer Research Center

Biomarkers that predict treatment effect heterogeneity across patient subgroups are important for treatment selection and have a potential to improve patient outcomes. A meta-analysis of individual participant data (IPD) is potentially more powerful than a single-study data analysis in evaluating markers for treatment selection. Our study was motivated by the IPD that were collected from three randomized control trials of hypertention and pre-eclampsia among

pregnant women to evaluate the effect of labor induction over expectant management of the pregnancy in preventing maternal complications. The existing literature on statistical methods for biomarker evaluation in IPD meta-analysis have evaluated a marker's performance in terms of its ability to predict risk of disease outcome, which do not directly apply to the treatment selection problem. In this study, we propose a statistical framework for evaluating markers for treatment selection given IPD meta-analytic data from small number of individual studies. We derive marker-based treatment rules by minimizing the average expected outcome across studies. We measure the performance of the marker-based rules by aggregating the study-specific performance, which is a clinically relevant and interpretable. Simulation studies and the application of the proposed methods to IPD meta-analytic data from the high risk pregnancy study are presented.

e-mail: *crkang@pitt.edu*

COMPARING THE SURROGACY OF MULTIPLE VACCINE-INDUCED IMMUNE RESPONSE BIOMARKERS IN HIV PREVENTION

Sayan Dasgupta*, Fred Hutchinson Cancer Research Center

Ying Huang, Fred Hutchinson Cancer Research Center

Identifying biomarkers to be used as surrogates for clinical end-points in randomized trials is useful for reducing study-periods and costs, relieving participants of unnecessary discomfort, and understanding treatment-effect mechanism. Here, for studying a vaccine's effect through immune responses in preventing HIV infection, we propose to use risk models with multiple vaccine-induced immune response biomarkers to measure the causal association between vaccine's effects on these biomarkers with that on the clinical end-point. In this setup, our main question of interest is to quantify these surrogate values to evaluate the effectiveness of the biomarkers in a comparative format, that is, to select markers with high surrogacy from a list of many, and hence save on cost and computation of the rest. Moreover this way we end up with a more parsimonious model which can potentially increase the predictive quality of the true markers. To address missing potential biomarker value if receiving vaccine, we make use of baseline predictors of the immune-response biomarkers and the augmented trial design (Follmann 1996). We then effectively make use of model assumptions to impute the missing marker values and conduct marker selection through a stepwise resampling and imputation method called stability selection (Long and Johson 2015).

e-mail: *sdasgup2@fredhutch.org*

COMPARISON OF METHODS FOR UPDATING RISK PREDICTION MODELS

Sonja Grill*, Technical University Munich, Germany

Donna P. Ankerst, Technical University Munich, Germany and University of Texas Health Science Center, San Antonio

Ruth M. Pfeiffer, National Cancer Institute, National Institutes of Health

As molecular markers become available, it is desirable to update existing risk prediction models with this new information. We thus assessed the performance of seven approaches for updating risk models, including a constrained maximum likelihood (CML) method (Chatterjee et al. 2015), an offset approach by Albert (1982), an approach allowing for a shrinkage parameter by Spiegelhalter/Knill-Jones (SKJ, 1984), a clean-slate approach of refitting a model on the new data and three methods using the likelihood ratio (LR) of the new marker, one assuming independence and two allowing dependence between the predictors by fitting two or one joint model to cases and controls. We studied the impact of the assumption of independence and ways to allow for dependence between a new marker and variables that are available on a new dataset. Using simulations, we assessed the bias by comparing observed and expected events in independent validation data and variability of predictions. Models updated using LR independent, SKJ and the Albert approach showed a large bias especially in scenarios with high dependence between predictors. The largest variability was observed for the clean-slate and the Albert approach. CML and the joint LR method performed consistently best.F217.

e-mail: *sonja.grill@tum.de*

EVALUATION OF BIOMARKER IDENTIFICATION THROUGH LIKELIHOOD RATIO TEST

Yu-Chuan Chen*, U.S. Food and Drug Administration

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

Precision medicine aims to apply molecular technologies and statistical methods to identify genomic biomarkers that indicate differential treatment responses so that each patient's treatment assignment can be optimized. Generally, the development of a biomarker model to classify patients for treatment assignment consists of two components: biomarker identification and subgroup selection. Biomarker identification plays an important role here since the performance a biomarker model depends on the identified biomarkers. We propose employing a likelihood ratio test (LRT) to test underlying subgroup structure defined by the selected biomarkers in order to determine if these biomarkers are truly helpful in clustering patients into two subgroups. Subgroup selection can be only carried out if the statistical test is significant. Additionally, we propose a new algorithm by applying 2-means clustering algorithm to divide patients into two subgroups. Binary classifiers are then developed using these two subgroups to make predictions on new patients.

e-mail: *yu-chuan.chen@fda.hhs.gov*

META-ANALYSIS OF PREDICTIVE VALUES OF BIOMARKERS

Mun Sang Yue*, Brown University

Constantine A. Gatsonis, Brown University

Methodological developments in the meta-analysis of studies of the diagnostic accuracy of tests, as measured by sensitivity and specificity, have been extensive and rich. In contrast, methods for the meta-analysis of studies of the predictive accuracy of tests, for example as measured by the positive and negative predictive value (PPV and NPV respectively), have received less attention despite the clinical relevance of prediction. As biomarkers become an increasingly essential tool in clinical medicine, methodological developments in the meta-analysis of the predictive accuracy of tests are needed. In this study, we propose a hierarchical summary predictive ROC (HSPROC) curve model to summarize estimates of PPV, NPV and disease prevalence jointly. The model accounts for the relation between PPV and NPV stemming from the dependence on the threshold for test positivity, and also addresses the monotonicity of the predictive ROC (PROC) curve. The HSPROC curves generated from the model can be used for comparison of different biomarkers. We applied the proposed method to a meta-analysis of prognostic capabilities of biomarkers for rapid rule-out of acute myocardial infarction, and compared the prognostic capabilities of these biomarkers.

e-mail: *mun_sang_yue@brown.edu*

116. COMPETING RISKS

CHECKING FINE AND GRAY MODEL WITH CUMULATIVE SUMS OF RESIDUALS: THEORY AND IMPLEMENTATION

Jianing Li*, Merck

Mei-Jie Zhang, Medical College of Wisconsin

Thomas H. Scheike, University of Copenhagen

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 paper, we present a class of analytical methods and graphical approaches for checking the assumptions of the Fine and Gray model. The proposed goodness-of-fit testing procedures are based on the cumulative sums of residuals, which validate the model in three aspects: (1) proportionality of hazard ratio, (2) linear functional form and (3) link function. For each assumption, we provide ap-value and a visualized plot against the null hypothesis using a simulation-based approach. The R-package crskdiagthat implements the proposed approaches has been released for public use, and it will be illustrated with two real data examples.

e-mail: *kinger198549@gmail.com*

COMPETING RISKS MODEL OF SCREENING AND SYMPTOMS DIAGNOSIS

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 semiparametrically using estimating equations and the NPMLE method. Asymptotic properties of the proposed estimators are established. The methodology is illustrated by simulation studies and applications using data from European cancer registries (EUREG).

e-mail: *shqiu@umich.edu*

CAUSE-SPECIFIC HAZARD REGRESSION FOR COMPETING RISKS DATA UNDER INTERVAL CENSORING AND LEFT TRUNCATION

Chenxi Li*, Michigan State University

Inference for cause-specific hazards from competing risks data under interval censoring and possible left truncation has been understudied. Aiming at this target, we develop a penalized likelihood approach for a Cox-type proportional cause-specific hazards model. The associated asymptotic theory is discussed. Monte Carlo simulations show that the approach performs very well for moderate sample sizes. We apply the method to event time data from a longitudinal study of dementia. The age-specific hazards of dementia and death are estimated, and risk factors of both are studied.

e-mail: *cli@epi.msu.edu*

EVALUATING CENTER PERFORMANCE ON COMPETING OUTCOMES

Sai Hurrish Dharmarajan*, University of Michigan

Douglas E. Schaubel, University of Michigan

It is often of interest to compare centers or healthcare providers on quality of care delivered. We consider the setting where evaluation of center performance on multiple competing events is of interest. We propose estimating center effects through cause-specific proportional hazards frailty models that allow for the correlation of a center's cause-specific effects. Estimation of our model proceeds

via penalized partial likelihood and is implemented in R. To evaluate center performances we also propose a direct standardized excess cumulative incidence (ECI) measure. For a given center, the ECI is defined as the difference between estimated cause-specific cumulative incidence functions (CIFs) for the entire study population at that center and the estimated CIF for the entire study population at an average center. Through simulations, we compare the proposed method to methods estimating center performance through cause-specific proportional hazard models assuming either (i) uncorrelated random center effects or (ii) fixed center effects, and demonstrate the advantages of using the proposed method to detect outlying centers. Using data from the Scientific Registry of Transplant Recipients, we apply our method to evaluate the performance of Organ Procurement Organizations on two competing risks: (i) receipt of a transplant and (ii) death on wait-list.

e-mail: *shdharma@umich.edu*

ADAPTIVE GROUP BRIDGE FOR COMPETING RISKS DATA

Natasha A. Sahr*, Medical College of Wisconsin

Kwang Woo Ahn, Medical College of Wisconsin

Anjishnu Banerjee, Medical College of Wisconsin

Competing risk regression analysis for clustered survival data with a large number of covariates is a challenging and relatively unexplored research area. In this context, we propose an adaptive group bridge penalty to select variables for proportional subdistribution hazards models with competing risks data. The proposed method selects not only group variables, but also within-group variables. The adaptive group bridge method was compared to the group bridge penalty and best subsets regression using AIC, BIC, and BICc criterion. The adaptive group bridge method was shown to correctly identify true individual covariates, that is, within group variables, better than the other methods in the simulation study. In addition, the adaptive group bridge was proficient in correctly selecting the true groups. Also, this method reduces bias when compared to the group bridge.

e-mail: *nsahr@mcw.edu*

EVALUATING UTILITY MEASUREMENT FROM RECURRENT MARKER PROCESSES IN THE PRESENCE OF COMPETING TERMINAL EVENTS

Yifei Sun*, Johns Hopkins University

Mei-Cheng Wang, Johns Hopkins University

In follow-up or surveillance studies, marker data are frequently collected conditioning on the occurrence of recurrent events. In many situations, a marker measurement does not exist unless a recurrent event takes place. Examples include medical cost for inpatient or outpatient cares, length-of-stay for hospitalizations, and prognostic or quality-of-life measurement repeatedly measured at multiple infections related to a certain disease. A recurrent marker process, defined between a pre-specified time origin and a terminal event, is composed of recurrent events and repeatedly measured marker measurements. This paper considers nonparametric estimation of the mean recurrent marker process in the situation where the occurrence of terminal event is subject to competing risks. Statistical methods and inference are developed to address a variety of questions and applications, for the purposes of estimating and comparing the integrated risk in relation to recurrent events, marker measurements and time to the terminal event for different competing risk groups. Furthermore, we develop an estimating procedure with improved efficiency. Simulation studies are conducted to evaluate the finite sample properties of the proposed estimators. The methods are applied to medical cost data for illustration.

e-mail: *ysun26@jhu.edu*

117. GWAS: TESTING

MEASURING AND TESTING DEPENDENCE BY KERNELIZED RV COEFFICIENT

Xiang Zhan*, Fred Hutchinson Cancer Research Center

Ni Zhao, Fred Hutchinson Cancer Research Center

Michael C. Wu, Fred Hutchinson Cancer Research Center

We propose a kernelized RV (KRV) coefficient as a new measure of dependence between two random vectors. Unlike the classical definition of RV coefficient, KRV is zero if and only if the random vectors are independent. A empirical KRV coefficient is also proposed, and is shown to be a consistent estimator of the population KRV coefficient. Finally, we study the distribution of the empirical KRV coefficient under the null hypothesis that two random vectors are independent. We build an independence test based on the empirical KRV coefficient and its null distribution. Simulation studies show that our KRV test has superior performance than some existing alternatives.

e-mail: *xiangzhan9@gmail.com*

ADAPTIVE GENE- AND PATHWAY-TRAIT ASSOCIATION TESTING WITH GWAS SUMMARY STATISTICS

Il-Youp Kwak*, University of Minnesota

Wei Pan, University of Minnesota

Pan et al, (2014) proposed a data-adaptive (aSPU) approach based on estimating and selecting the most powerful test among a class of sum of powered score (SPU) tests, which cover several popular

tests as special cases. Furthermore, Pan et al, (2015) extended the methodology to pathway analysis (aSPUpath). In particular, two parameters are introduced such that the test is adaptive at both the SNP and gene levels. However, the two adaptive tests for gene- and pathway-trait associations are only applicable to the case with individual-level genotype and phenotype data. It is often difficult to obtain access to individual-level data. We propose extending the two adaptive tests to the case with only summary statistics for individual SNPs, demonstrating their applications to a meta-analyzed GWAS dataset. The methods are available in R package, aSPU.

e-mail: *ikwak@umn.edu*

SNP-SET TESTS USING GENERALIZED BERK-JONES STATISTICS IN GENETIC ASSOCIATION STUDIES

Ryan Sun*, Harvard University

Xihong Lin, Harvard University

It is often of interest to test whether a group of SNPs is associated with certain traits or diseases. For instance, natural groupings of SNPs may arise from their locations in a common gene or pathway of interest. However, existing methodologies to perform these tests may be subject to power loss due to the fact that the associations between SNPs and an outcome are generally sparse and weak. Motivated by the Berk-Jones (BJ) statistic, which is notable for its asymptotic ability to detect sparse and weak signals, we propose a new test for association between a SNP-set and an outcome - the generalized Berk-Jones (gBJ) statistic. The standard Berk-Jones statistic is unsuited for genetic association studies because it assumes SNPs within a set, e.g., a gene, are independent when calculating p-values. Our proposed generalized Berk-Jones statistics allow for an arbitrary correlation structure among SNPs, and gBJ is able to perform an accurate analytical p-value calculation accounting for correlation. We compare the performance of mBJ to other SNP-set tests across a range of genetic architectures, varying the signal strength and sparsity of causal SNPs. We also apply the methods to analyze data from an infant neurodevelopment GWAS.

e-mail: *ryanrsun@gmail.com*

TESTING FOR GENETIC ASSOCIATIONS IN ARBITRARILY STRUCTURED POPULATIONS

Minsun Song*, University of Nevada Reno

Wei Hao, Princeton University

John D. Storey, Princeton University

We present a new statistical test of association between a trait and genetic markers, which we theoretically and practically prove to be robust to arbitrarily complex population structure. The statistical test involves a set of parameters that can be directly estimated from large-scale genotyping data, such as that measured in genome-wide associations studies (GWAS). We also derive a new set of methodologies, called a genotype-conditional association test (GCAT), shown to provide accurate association tests in populations with complex structures, manifested in both the genetic and environmental contributions to the trait. Our proposed framework provides a substantially different approach to the problem from existing methods.

e-mail: *minsuns@unr.edu*

NOVEL STATISTICAL TEST FOR GENETIC PLEIOTROPY

Daniel J. Schaid*, Mayo Clinic

The statistical association of a single trait with genetic data has revolutionized human genetics, with many genome-wide association studies providing guidance on genetic factors influencing human health. Pleiotropy – the association of more than one trait with a genetic marker – is believed to be common, yet current multivariate methods do not formally test pleiotropy. Current multivariate methods, such as multivariate regression of multiple traits on a genetic marker, or reverse regression of a genetic marker on multiple traits, test the null hypothesis that no traits are associated with a genetic marker; a statistically significant finding could result from only one trait driving the association. A formal null hypothesis of pleiotropy should allow none or one trait associated with a genetic marker, so that rejection of the null implies at least two traits are associated with the marker. We developed a new likelihood ratio statistic for quantitative traits to test the null hypothesis of pleiotropy. The new methods, with simulations illustrating it properties, will be presented, as well as application to a study of the genetics of response to small pox vaccination.

e-mail: *schaid.daniel@mayo.edu*

AN ADAPTIVE MULTIVARIATE TEST IN APPLICATIONS TO MULTIPLE TRAIT-MULTIPLE GENETIC VARIANT ASSOCIATIONS FOR GWAS AND SEQUENCE DATA

Junghi Kim*, University of Minnesota

Wei Pan, University of Minnesota

Testing for genetic association with multivariate trait has become increasingly important, not only because of its potential to boost statistical power, but also for its direct relevance to some applications. For example, there is accumulating evidence showing that some complex neurodegenerative and psychiatric diseases like Alzheimer's are due to disrupted brain networks, for which it would be natural to identify genetic variants associated with the disrupted brain networks. In spite of its promise, testing for multivariate trait associations is challenging: if not appropriately used, its power can

be much lower than testing on each univariate trait separately (with a proper control for multiple testing). Furthermore, differing from most existing methods for single SNP-multiple trait associations, we consider gene-based association testing for multiple traits, due to well known genetic heterogeneity and small effect sizes of individual SNPs. Because the power of a test critically depends on several unknown factors such as the proportions of associated SNPs, genes and traits among those tested, we propose a flexible test that data- adaptively determines some optimal parameters in the test to yield high power across a wide spectrum of situations. We compare the performance of the new test with existing tests using both simulated and real data. The proposed test was applied to structural MRI data drawn from the Alzheimer's Disease Neuroimaging Initiative (ADNI) project to identify genetic variants associated with the human brain default mode network (DMN).

e-mail: *kimx2859@umn.edu*

118. MEASUREMENT ERROR

THE ORTHOGONALLY PARTITIONED EM ALGORITHM: EXTENDING THE EM ALGORITHM FOR ALGORITHMIC STABILITY FOR BIAS CORRECTION DUE TO IMPERFECT DATA

Michael Regier*, West Virginia University

Erica Moodie, McGill University

We propose an extension of the EM algorithm that exploits the assumption of unique parameterization, reduces bias due to imperfect data (measurement error and/or missing data), converges when standard implementation of the EM algorithm has a low convergence probability, and reduces a potentially complex algorithm into a sequence of smaller, simpler, self-contained EM algorithms. Both theoretical and finite sample results are considered. Evidence suggests that partitioning can provide better bias reduction in the estimation of model parameters. Breaking down a complex problem in to simpler steps may make the EM algorithm more accessible to a wider and more general audience permitting a broader implementation of the EM algorithm.

e-mail: *mregier@hsc.wvu.edu*

A SIMULATION STUDY OF NONPARAMETRIC TOTAL DEVIATION INDEX AS A MEASURE OF AGREEMENT BASED ON QUANTILE REGRESSION

Yi Pan*, Centers for Disease Control and Prevention

Lawrence Lin, JBS Consulting Services Company

A.S. Hedayat, University of Illinois, Chicago

Huiman Barnhart, Duke Clinical Research Institute, Duke University

Michael Haber, Emory University

Total deviation index (TDI) captures a pre-specified quantile of the absolute deviation of paired observations from raters, observers, methods, assays, instruments, etc. We compare the performance of total deviation index (TDI) using nonparametric quantile regression to the TDI assuming normality (Lin 2000). This simulation study considers 3 distributions: normal, Poisson, and uniform at quantile levels of 0.8 and 0.9 for cases with and without contamination. Study endpoints include the bias of TDI estimates (compared to their respective theoretical values), standard error of TDI estimates (compared to their true simulated standard errors), and test size (compared to 0.05) and power. Nonparametric TDI using quantile regression, although it slightly underestimates and delivers slightly less power for data without contamination, works satisfactorily under all simulated cases even for moderate (say, 40) sample sizes. The performance of the TDI based on a quantile of 0.8 is in general superior to that of 0.9. The performances of nonparametric and parametric TDI methods are compared with a real data example. Nonparametric TDI can be very useful when the underlying distribution on the difference is not normal, especially when it has a heavy tail.

e-mail: *jnu5@cdc.gov*

THERE IS NO IMPACT OF EXPOSURE MEASUREMENT ERROR ON LATENCY ESTIMATION IN LINEAR MODELS

Sarah B. Peskoe*, Harvard School of Public Health

Molin Wang, Harvard School of Public Health

Donna Spiegelman, Harvard School of Public Health

Identification of the latency period of a time-varying exposure is key when assessing many environmental, nutritional, and behavioral risk factors. A pre-specified exposure metric involving an unknown latency parameter is often used in the statistical model for the exposure-disease relationship. Likelihood-based methods have been developed to estimate this latency parameter for generalized linear models, but are nonexistent for scenarios where the exposure is measured with error, as is usually the case. Here, we explore the performance of naïve estimators for both the latency parameter and the regression coefficients, which ignore exposure measurement error, assuming a linear measurement error model. We prove that in many scenarios under this general measurement error setting, the maximum likelihood estimator for the latency parameter remains consistent, while the regression coefficient estimates are inconsistent as previously obtained under standard measurement error models where the primary disease model does not involve a latency parameter. Conditions under which this result

holds will be generalized to the extent possible, and simulations will be presented, exploring the applicability of these findings to Cox models and non-linear measurement error models. The findings are illustrated in a study of body mass index in relation to cigarette smoking in the Nurses' Health Study.

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

IMPROVED ESTIMATION FOR HIGH DIMENSIONAL MEASUREMENT ERROR MODELS

Abhishek Kaul*, National Institute of Environmental Health Sciences, National Institutes of Health

In this paper we investigate estimation via bias corrected least squares post model selection in the context of high dimension measurement error models. We show that by separating model selection and estimation, it is possible to achieve improved convergence rates of the l2 estimation error in comparison to simultaneous estimation and variable selection methods such as l1 penalized likelihood, i.e., faster than sqrt{slogp/n}. In fact under perfect (w.p.? 1) model selection, the l2 rate of convergence is indeed the oracle rate of sqrt{s/n}. Here s, p are the number of non zero parameters and the model dimension respectively, n is the sample size. We show that under very general model selection criteria, the proposed method provides three major advantages, first, it is at least as efficient as l1 penalized method. Second, performing model selection without the availability of the covariate noise covariance matrix. Lastly, being able to provide estimates with only a small sub-block of this covariance matrix. Furthermore we also show that the model selection requirements are met by the SIS method of Fan and Lv (2008). All results are supported empirically by a simulation study.

e-mail: *abhishek.kaul@nih.gov*

IDENTIFYING HEAT WAVES IN FLORIDA: THE IMPACT OF MISSING EXPOSURE DATA AND THRESHOLDS ON MISSINGNESS

Emily Leary*, University of Missouri, Columbia

Linda J. Young, University of Florida

Background: Using current climate models, regional-scale changes for Florida over the next 100 years are predicted to include warming over terrestrial areas and very likely increases in the number of high temperature extremes. No uniform definition of a heat wave exists and most heat wave definitions used in public health do not consider impacts of missing weather (exposure) information. Objectives: To identify and describe methods of imputing missing weather data and how those methods can affect identified periods of extreme heat in Florida. In addition, thresholds for where the missing data starts to influence the overall results will be identified. Methods: In addition to ignoring missing data, temporal, spatial, and spatio-temporal models are described and utilized to impute missing historical weather data from 1973 to 2012 from 43 Florida weather monitors. Calculated thresholds are used to define periods of extreme heat across Florida and ongoing simulations will be used to determine thresholds for missing data effects. Results: Modeling of missing data and imputing missing values can affect the identified periods of extreme heat, through the missing data itself or through the computed definitions.

e-mail: *learye@health.missouri.edu*

OPTIMAL DESIGN STRATEGY TO ACHIEVE A PRE-SPECIFIED POWER WHEN THE BIOMARKER IS SUBJECT TO MEASUREMENT ERROR

Matthew T. White*, Boston Children's Hospital

Sharon X. Xie, University of Pennsylvania

The discovery of effective diagnostic biomarkers is an important and highly active area of research. This discovery is often impeded, however, when biomarkers are obtained with measurement error, which may cause the biomarker to appear ineffective if not taken into account in the analysis. We develop an optimal design strategy to study the effectiveness of an error-prone biomarker in differentiating diseased from non-diseased individuals and focus on the area under the receiver operating characteristic curve (AUC) as the primary measure of effectiveness. Using an internal reliability sample within the diseased and non-diseased groups, we develop an optimal study design strategy that achieves a pre-specified power. We develop optimal allocations of the number of subjects, the size of the reliability sample, and the number of replicate observations per subject in the reliability sample within each group under a variety of commonly-seen study conditions and show that the pre-specified power is achieved through extensive simulations.

e-mail: *matthew.white@childrens.harvard.edu*

THE ESTIMATION OF MISCLASSIFICATION VIA CONTINUOUS-TIME HIDDEN MARKOV MODEL

Liqiong Fan*, Medical University of South Carolina

Sharon Yeatts, Medical University of South Carolina

Background: Although transitions between disease states are continuous, subject evaluations are usually panel-observed, resulting in discrete characterizations which may be subject to misclassification. Applications of the continuous time Hidden Markov Model (CTHMM) for misclassified disease outcomes have been proposed. However, model performance, in terms of estimation and model diagnosis, is not well described. We provide a simulation study to demonstrate the performance of the model with a focus on misclassification. Methods: Data were simulated with pre-specified transition intensi-

ties and misclassification probabilities with varying state space, number of observations per subject, and sample size. Results based on both basic Markov Model (MM) and CTHMM were compared. Results: In the two-state CTHMM, the misclassification probabilities are well estimated for mild or moderate misclassification but slightly underestimated for severe misclassification. For a three-state CTHMM, more bias was observed with moderate misclassification compared to the two-state case. The Pearson-type goodness-of-fit test cannot well distinguish between CTHMM and MM for misclassified data. Conclusion: CTHMM can be used to estimate misclassification rate in the data with relatively small amount of misclassification. However, further research is needed to better identify the necessity of the latent structure.

e-mail: *fanliq@musc.edu*

119. STATISTICAL GENETICS

A FUNCTIONAL WEIGHTED U TEST FOR DETECTING GENE-GENE INTERACTIONS

Pei Geng*, Michigan State University

Qing Lu, Michigan State University

This article proposes a functional weighted U (FWU) for identifying gene-gene interactions associated with a disease phenotype. In FWU, a functional approach is used to capture linkage disequilibrium among multiple sequencing variants on a gene. Based on the smoothing curve of each subject fitted by the functional approach, L-2 norm distance is used to measure the genetic similarity between two subjects. A functional weighted U statistic is then defined based on the product of the genetic similarity and the phenotypic similarity. Under the null hypotheses of no association between genes and the phenotype, it is proved that FWU is asymptotically normally distributed. Through simulations, we investigated the type I error and the power of the new method, and compared it with several existing methods. Finally, we applied FWU to a large-scale substance dependence (SD) sequencing data, evaluating gene-gene interactions among SD-related genes.

e-mail: *gengpei@msu.edu*

TIME-COURSE GENE SET ANALYSIS OF LONGITUDINAL RNA-Seq DATA

Boris P. Hejblum*, Harvard School of Public Health

Denis M. Agniel, Harvard Medical School

As gene expression measurement technology is shifting from microarrays to sequencing, statistical tools derived for the subsequent data analysis must be adapted. Since RNA-seq data are measured as counts, methods developed for microarrays are unsuitable without modification. Recently, it has been proposed to tackle the count nature of these data by modeling gene read log-counts as continuous variables using precision weights to account for heteroscedasticity. We adopt this approach and propose an efficient top-down method to detect longitudinal changes in RNA-seq gene sets. Usinga prioridefined gene sets, we identify those whose expression varies over time with a variance component score test that accounts both for covariates and precision weights (which may be estimated from a nonparametric model). We demonstrate that, despite the presence of the nonparametric weights, the test statistic has a simple form and limiting distribution, both of which may be computed quickly. The proposed method is applied to both simulated data and a real dataset.

e-mail: *bhejblum@hsph.harvard.edu*

A POWERFUL AND DATA-ADAPTIVE TEST FOR RARE VARIANT-BASED GXE ANALYSIS

Tianzhong Yang*, University of Texas School of Public Health

Peng Wei, University of Texas School of Public Health

As whole exome/genome sequencing data become increasingly available in large genetic epidemiology research consortia, there is an emerging interest in testing the interaction between rare genetic variants and environmental exposures. However, testing rare variant-based GxE is more challenging than testing genetic main effects due to the difficulty in estimating the latter under the null hypothesis of no GxE effects and the presence of neural variants, which are ubiquitous in sequencing data. In response, we developed a powerful and data-adaptive GxE test, called “aGE”, in the framework of aSPU test (Pan et al, AJHG 2015), originally proposed for testing the main effects of rare variants. Using extensive simulations, we compared our proposed test with rareGE (Chen et al, Hum Here 2014) and iSKAT (Lin et al, Biometrics 2015). We found that the proposed test could control the Type I error in the presence of a large number of neutral variants, whereas the iSKAT method could suffer from inflated Type I error. In addition, our test was more resilient to inclusion of neutral variants and more powerful than the rareGE method. Finally, we demonstrate the performance of the proposed “aGE” test using the UK10K sequencing data.

e-mail: *tianzhong.yang@uth.tmc.edu*

MEDIATION METHODS FOR CASE-CONTROL SETTINGS WITH APPLICATIONS TO GENOMICS

Sheila M. Gaynor*, Harvard School of Public Health

Xihong Lin, Harvard School of Public Health

Mediation methods have been developed to characterize causal relationships that act through a mediator. Traditional approaches are easily applied to observational data and oftentimes can be adapted

to case-control data. However, there are many settings in which the modeling assumptions of these methods, such as rarity of a binary outcome, are not appropriate. This setting is present in genomic data, particularly when analyzing common tumor subtypes within a cancer type. Recent studies have suggested that genomic analyses may be improved by jointly analyzing SNP and gene expression data to analyze phenotypes of complex diseases. By developing mediation methods for case-control settings with common outcomes, we are able to jointly analyze genetic and genomic data for complex diseases. We propose a method for mediation in case-control studies with a common binary outcome. The proposed method uses generalized linear regression and incorporates weights to adjust for case status. The method is assessed through simulations in both low-dimensional and high-dimensional settings. Further, we demonstrate that our method can be used to identify effects of interest in cancer samples from The Cancer Genome Atlas.

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

A RANDOM FIELD METHOD FOR GENETIC ASSOCIATION ANALYSIS OF CORRELATED PHENOTYPES DERIVED FROM ELECTRONIC MEDICAL RECORDS

Xue Zhong*, Vanderbilt University
Nancy J. Cox, Vanderbilt University

Electronic medical records (EMRs) are increasingly employed for genetics studies to infer the relationship between genetics and disease. In a phenome-wide association study of EMRs, a typical analysis approach involves univariate tests between genotype-phenotype pairs followed by Bonferroni correction. This strategy ignores correlations among the phenotypes and over-corrects for the number of tests, thus resulting in loss of power. Here, we propose a random field method to analyze the related phenotypes together to increase the statistical power. This method utilizes prior knowledge on disease comorbidity pattern and models the phenotype correlation structure to identify genotype-phenotype associations. This method requires only summary statistics from standard univariate analyses and do not rely on individual-level genotype-phenotype data that are often restricted for availability. Other advantages include the ability to handle both continuous and binary phenotypes and to efficiently analyze thousands of phenotypes. We present simulation results and illustrate the applicability of the method by applying it to a DNA bioBank linked to extensive EMRs.

e-mail: *xue.zhong@vanderbilt.edu*

STATISTICAL CONSIDERATIONS IN ANALYTICAL VALIDATIONS FOR SEQUENCING BASED GENETIC TESTS

Jincao Wu*, U.S. Food and Drug Administration
Meijuan Li, U.S. Food and Drug Administration

With recent rapid development in molecular and genetic technology, the applications of genetic testing span medical disciplines, including: newborn screening for highly penetrant disorders; diagnostic and carrier testing for inherited disorders; risk prediction testing for complex disorders; and pharmacogenetic testing to guide individual therapeutic management. More recently, sequencing of the human genome creates the possibility for scientists to develop tools to transform the personalized medicine from an idea to a practice. However, despite extraordinary advances that have been made to date in medical fields, major analytical and interpretative challenges have emerged, ranging from the validation of large numbers of genetic variation in a patient, to managing the huge amount of data that accompany a single sequenced genome. In this talk, we will discuss the statistical issues and explore the regulatory pathways in analytical validation studies for sequencing based- genetic tests.

e-mail: *jincao.wu@fda.hhs.gov*

120. INFERENCE FOR BRAIN NETWORKS

POPULATION INFERENCE FOR FUNCTIONAL BRAIN CONNECTIVITY

Manjari Narayan, Rice University

Genevera I. Allen*, Rice University and Baylor College of Medicine
Many are interested in understanding how the brain communicates at a systems level and how these systems are disrupted in neurological conditions and diseases. By modeling the brain as a network of functional brain connections, we seek to find network patterns that are altered in a group of patients or are associated with symptom severity. To achieve this, we introduce new two-level models using Gaussian Graphical Models to describe subject-level networks and Generalized Linear Models to describe population-level effects as a function of subject-level network patterns. This problem leads to a new statistical paradigm which we term Population Post Selection Inference (popPSI). To address this, we present a new estimation and inference technique, R^3, which employs resampling, random penalization, and random effects test statistics. Our method offers substantial improvements in statistical powerful and yields fewer false positives than existing approaches. We use our techniques to discover alterations in functional brain networks of patients with autism and neurofibromatosis.

e-mail: *gallen@rice.edu*

ROBUST BRAIN STRUCTURAL CONNECTIVITY ANALYSIS USING HCP DATA

Zhengwu Zhang*, SAMSI and University of North Carolina, Chapel Hill
Antonio Canale, University of Turin
David B. Dunson, Duke University

Structural connection in an individual brain plays a fundamental role in how the mind responds to everyday tasks and challenges. Modern imaging technology such as diffusion MRI (dMRI) makes it easy to peer into an individual brain and collect valuable data to infer the structural connectivity. The difficulty for current statistical analysis and inference of the connectivity of human brain is to extract robust and high-resolution connectivity network from dMRI. In this talk, I will introduce the Human Connectome Project Dataset, which provides high quality structural and diffusion MRI. Moreover, I will present a state-of-the-art dMRI processing pipeline to process the HCP data to reliably construct the structural connectivity from dMRI. The pipeline includes streamline construction, streamline compression, and robust connectivity coupling strength extraction.

e-mail: *zhengwu@stat.fsu.edu*

NODE-WISE INFERENCE FOR GROUPS OF CONNECTIVITY GRAPHS

Philip T. Reiss*, New York University School of Medicine

Group analyses of functional connectivity typically seek differences in brain connectivity patterns between groups defined by demographic variables or by diagnosis. This talk will present a recently proposed method that applies distance-based permutation tests for group differences in node-specific connectivity patterns' i.e., in rows of a matrix of connectivity scores, such as correlations, among a set of regions of interest. This node-by-node testing poses a less severe multiplicity problem than edge-by-edge testing, and thus can detect effects that the latter may miss. A key disadvantage of node-wise testing, however, is interpretability: it identifies regions (nodes) with between-group connectivity differences, but does not describe these differences in terms of brain networks or circuits. We propose novel follow-up analyses that can partially overcome this limitation. To illustrate these ideas, we examine functional connectivity abnormalities associated with psychopathology, using data from a large neurodevelopmental cohort study.

e-mail: *phil.reiss@nyumc.org*

DISENTANGLING BRAIN GRAPHS: THE CONFLATION OF NETWORK AND CONNECTIVITY INFERENCE

Sean L. Simpson*, Wake Forest School of Medicine

Paul J. Laurienti, Wake Forest School of Medicine

Understanding the human brain remains the Holy Grail in biomedical science, and arguably in all of the sciences. Our brains represent the most complex systems in the world (and some contend the universe) comprising nearly one hundred billion neurons with septillions of possible connections between them. The structure of these connections engenders an efficient hierarchical system capable of consciousness, as well as complex thoughts, feelings, and behaviors. Brain connectivity and network analyses have exploded over the last decade due to their potential in helping us understand both normal and abnormal brain function. Functional connectivity (FC) analysis examines functional associations between time series pairs in specified brain voxels or regions. Brain network analysis serves as a distinct subfield of connectivity analysis in which associations are quantified for all time series pairs to create an interconnected representation of the brain (a brain network), which allows studying its systemic properties. While connectivity analyses underlie network analyses, the subtle distinction between the two research areas has generally been overlooked in the literature, with them often being referred to synonymously. However, developing more useful analytic methods and allowing for more precise biological interpretations requires distinguishing these two complementary domains. Here we briefly delineate methods for connectivity and network inference and discuss the importance of joint and hybrid methodology for expanding the scope of neuroscience research.

e-mail: *slsimpso@wakehealth.edu*

121. RECENT DEVELOPMENT IN JOINT MODELING FOR LONGITUDINAL DATA

BAYESIAN METHODS FOR NON-IGNORABLE DROPOUT IN JOINT MODELS IN SMOKING CESSATION STUDIES

Jeremy Gaskins, University of Louisville

Michael J. Daniels*, University of Texas, Austin

Inference on data with missingness can be challenging, particularly if the knowledge that a measurement was unobserved provides information about its distribution. Our work is motivated by the Commit to Quit II study, a smoking cessation trial that measured smoking status and weight change as weekly outcomes. It is expected that dropout in this study was informative and that patients with missed measurements are more likely to be smoking, even after conditioning on their observed smoking and weight history. We jointly model the categorical smoking status and continuous weight change outcomes by assuming normal latent variables for cessation and by extending the usual pattern mixture model to the bivariate case. The model includes a novel approach to sharing information across patterns through a Bayesian shrinkage framework to improve estimation stability for sparsely observed patterns. To accommodate the presumed informativeness of the missing data in a parsimonious manner, we model the unidentified components of the model under a non-future dependence assumption and specify departures from missing at random through sensitivity parameters, whose distributions are elicited from a subject-matter expert.

e-mail: *mjdaniels@austin.utexas.edu*

MEAN-CORRELATION REGRESSION FOR DISCRETE LONGITUDINAL RESPONSES

Cheng Yong Tang*, Temple University

Weiping Zhang, University of Science and Technology of China

Chenlei Leng, University of Warwick

Joint mean-covariance regression modelling with unconstrained parameterization has provided statisticians and practitioners a powerful analytical device for characterising covariations between continuous longitudinal responses. How to develop a delineation of such an unconstrained regression framework amongst categorical or discrete longitudinal responses, however, remains an open and challenging problem. This paper studies, for the first time, a novel mean-correlation regression for a family of generic discrete responses. Targeting at the joint distributions of the discrete longitudinal responses, our regression approach is constructed by using an innovative copula model whose correlation parameters are represented by unconstrained hyperspherical coordinates. To overcome the computational intractability in maximising the full likelihood of the discrete responses in practice, we develop a computationally efficient pairwise likelihood approach for estimation. We show that the resulting estimators of the proposed approaches are consistent and asymptotically normal. A pairwise likelihood ratio test is further proposed for statistical inference. We demonstrate the effectiveness, parsimoniousness and desirable performance of the proposed approach by analysing three discrete longitudinal data sets and conducting extensive simulations.

e-mail: *yongtang@temple.edu*

SIMULTANEOUS MEAN AND COVARIANCE MODELING OF CHRONIC KIDNEY DISEASE

Xiaoyue Niu*, The Pennsylvania State University

Peter Hoff, University of Washington

Chronic kidney disease (CKD) is a serious health condition that is associated with premature mortality and decreased quality of life. Most of the existing population-level CKD studies focus on either the prevalence of CKD or the associations of CKD with other health problems. Very few studies have examined these quantities jointly, or have examined how the associations between CKD and other health problems vary by demographic characteristics. In this article, we propose a joint mean and covariance regression model to statistically describe how the quantitative measurement of CKD and associations between CKD and other health problems vary according to demographic predictors. We apply the methodology to the NHANES 2010 data and discuss guidelines for model selection and evaluation using standard criteria such as AIC in conjunction with posterior predictive goodness of fit plots. With the fitted results from the model, we are able to identify sub-populations that are at high risk of developing CKD, as well as those for which CKD is likely to co-occur with other health problems.

e-mail: *xiaoyue@psu.edu*

122. MODEL VALIDATION: ITS CONCEPT, STATISTICAL INTEGRITY, REGULATORY LANDSCAPE, AND INDUSTRY APPLICATION

ISSUES WITH TRAINING, TESTING AND VALIDATION DATASETS IN THE DEVELOPMENT OF DIAGNOSTIC DEVICES

R. Lakshmi Vishnuvajjala*, U.S. Food and Drug Administration

Model development and validation are critical parts in the development of classifiers, or diagnostics devices as they are called in submissions to FDA. The integrity of methods used to develop and validate classification models ensures the performance of the diagnostic devices in future subjects on whom the device is used There is a lot of confusion about training, testing and validation datasets, as well as internal and external validation of models. We will discuss some good practices for developing and validating classification models in diagnostic devices. We will also discuss some problems frequently encountered with training and validation datasets which can lead to overly optimistic estimates of performance metrics.

e-mail: *lakshmi.vishnuvajjala@fda.hhs.gov*

ESTABLISHING CLINICAL USEFULNESS OF A DIAGNOSTIC TEST INTENDED TO GUIDE THERAPY DECISIONS

Lisa M. McShane*, National Cancer Institute, National Institutes of Health

The term validation is used liberally in biomedical literature reporting studies of biomarkers and omics signatures or in vitro diagnostic tests based on them, but too infrequently do such claims establish clinical usefulness (clinical utility). Research teams aiming to develop in vitro diagnostic tests for the purpose of guiding therapy decisions must consider aspects of analytic validation, clinical validation, and clinical utility assessment which go far beyond determining whether results of diagnostic tests demonstrate a statistically significant association with outcome. It is important to establish early an intended clinical use and then plan a series of studies to evaluate analytical and clinical performance of the test in that context. Through a series of real examples, the importance of proper selection of study patients and specimens, clinical endpoints, and performance metrics, and use of appropriate statistical approaches are emphasized.

e-mail: *lm5h@nih.gov*

GUIDELINES FOR REPORTING STUDIES THAT DEVELOP OR VALIDATE A MULTIVARIABLE RISK PREDICTION MODEL

Doug Altman, University of Oxford

Gary Collins*, University of Oxford

Many reviews of published studies show that the methods and reporting of studies to develop or validate risk prediction models are often poor. Those studies cannot safely inform clinical practice. Poor reporting of methods hampers assessment of whether researchers used sound methodology. I will describe the recent TRIPOD guidelines for reporting studies risk of prediction models that develop and/or validate risk prediction models (TRIPOD). I will also present results of recent research on the methodology of validation studies for risk prediction models.

e-mail: *gary.collins@csm.ox.ac.uk*

MODEL VALIDATION: AN INDUSTRY CASE STUDY

Susan H. Gawel*, Abbott Labs

There are multiple approaches one could take in developing and validating models. In this presentation we will explore some of those options and review a case study explaining why a particular method was chosen over others.

e-mail: *susan.gawel@abbott.com*

123. MISSING DATA IN LONGITUDINAL CLINICAL TRIALS: CHOICE OF PRIMARY ESTIMAND AND ITS RELATIONSHIP TO THE STATISTICAL ANALYSIS METHODS

CHOICE OF ESTIMAND AND MISSING DATA IN CLINICAL TRIALS

Roderick J. Little*, University of Michigan

The choice of estimand has a major impact on methods for handling missing data in clinical trials, affecting issues such as the need to collect follow-up data after treatment discontinuation, the method for imputing missing data after discontinuation, and whether such imputations are needed in the first place. I discuss three alternative intention-to treat estimands: the average effect when individuals continue on the assigned treatment after discontinuation, the average effect when individuals take a control treatment after treatment discontinuation, and a summary measure of the effect of treatment prior to discontinuation. I argue that the latter choice of estimand has advantages and should receive more consideration. Ideas are illustrated on a past study of inhaled insulin treatments for diabetes, sponsored by Eli Lilly, and other examples. (Joint work with Shan Kang).

e-mail: *rlittle@umich.edu*

ROLE OF SIMULATIONS IN THE SELECTION OF THE PRIMARY ESTIMAND AND STATISTICAL METHODS FOR HANDLING MISSING DATA IN LONGITUDINAL TRIALS

Elena Polverejan*, Janssen R&D

Selection of the primary estimand and corresponding analysis methods in the presence of missing response information is a complex problem in longitudinal trials. This presentation uses a trial example for a chronic pain development program to describe how simulations can be a powerful tool to address this complex problem at the trial design stage. The 2014 FDA draft guidance for analgesic indications emphasizes the importance to attribute "bad outcomes" to subjects who were unable to complete the course of treatment because such subjects did not benefit from the treatment and recommends the utilization of methods that "avoid attributing an overall benefit to a drug that does not benefit individual subjects". The performed simulations incorporate various assumptions for the trial including the amount and reasons of missing data over time. Simulation metrics such as the estimated power and Type I error rate, mean and standard error for the treatment difference versus control, etc. allow a quantitative comparison of the performance of different methods that "penalize" some categories of missing data. The talk will show how the simulation results play a critical role towards an informed selection of the primary estimand, primary analysis method and additional sensitivity analyses.

e-mail: *EPolvere@its.jnj.com*

CHOOSING ESTIMANDS IN CLINICAL TRIALS WITH MISSING DATA

Craig H. Mallinckrodt*, Eli Lilly and Company

Recent research has fostered new guidance on preventing and treating missing data. Consensus exists that clear objectives should be defined along with the causal estimands, trial design and conduct should maximize adherence to the protocol specified interventions, and a sensible primary analysis should be paired with plausible sensitivity analyses. An estimand is simply what is to be estimated. Two general categories of estimands are: effects of the drug as actually taken (de-facto) and effects of the drug if taken as directed (de-jure). De-jure and de-facto estimands each have strengths and limitations. An iterative process including objectives, estimands, design, analysis, and sensitivity analyses can be used to guide protocol development. Objectives should reflect the diverse needs of regulators, payers, prescribers, patients, care givers, sponsors, and other researchers. Although design and analysis considerations should not dictate choice of estimand, these considerations should not be ignored. For example, maximizing adherence reduces sensitivity to missing data assumptions for de-jure estimands, but may reduce generalizability of results for de-facto estimands if the methods used to maximize adherence in the

trial are not feasible in clinical practice. Both de-jure and de-facto estimands are often needed to understand drug benefit and de-jure estimands will often be the focus of safety evaluations. Newer approaches such as reference based controlled imputation provide useful options for analyses of de-facto estimands. A sequential testing approach starting with a de-jure estimand(s) followed by a de-facto estimand(s) may be useful in assessing drug benefit.

e-mail: *mallinckrodt_craig@lilly.com*

124. BAYESIAN ANALYSIS OF COMPLEX SURVEY DATA

CLUSTER LIKE YOU DO: WHEN TO AVOID TRADITIONAL CLUSTERING APPROACHES IN THE PRESENCE OF SPARSE DATA

Rebecca C. Steorts*, Duke University

Most commonly-used generative models for clustering implicitly assume that the number of data points in each cluster grows linearly with the number of data points in the dataset. For example, this is the case in finite mixture models, Dirichlet process mixture models, and Pitman-Yor process mixture models, and in fact, any infinitely-exchangeable clustering model makes this linear-growth assumption. For certain applications, however, this is unrealistic. For example, when performing entity resolution---i.e., identifying duplicate records in large, noisy databases--often the process generating each cluster is unrelated to the overall size of the dataset, and as a result, each cluster may contain only a few points, even though the dataset may be very large. We call this type of scenario a small clustering problem. To explore how well commonly-used clustering models would be able to cope with small clustering problems, we perform prior predictive checks using real and simulated data. Further, we introduce a new model designed for small clustering, and assess it as well.

e-mail: *beka@stat.duke.edu*

SPATIAL SMOOTHING OF COMPLEX SURVEY DATA FOR SMALL AREA ESTIMATION

Jon Wakefield*, University of Washington

Small area estimation (SAE) is an important endeavor since many agencies require estimates of health, education and environmental measures in order to plan and allocate resources and target interventions. Often SAE is based on complex survey data, with sampled units having an associated sampling weight that reflects the design used. Spatial modeling is usually carried out within a model-based, as opposed to a randomization, paradigm, but ignoring the weights can be dangerous, since bias can result. We describe approaches to spatial modeling that acknowledge the design, but use Bayesian hierarchical smoothing models to alleviate imprecise inference in small areas. Implementation is fast since it is based on integrated nested Laplace approximation. The techniques are demonstrated using data on under-5 mortality in regions of Tanzania.

e-mail: *jonno@uw.edu*

MULTILEVEL REGRESSION AND POSTSTRATIFICATION FOR SURVEY WEIGHTED INFERENCE

Yajuan Si*, University of Wisconsin, Madison

Andrew Gelman, Columbia University

Survey weighting adjusts for differences between the collected samples and the target population. However, classical weights have lots of problems. Extreme values of weights will cause high variability and blow up the estimates. In practice, weighting construction requires arbitrary choices about selection of weighting factors and interactions, pooling of weighting cells and weight trimming. The general principles of Bayesian analysis imply that models for survey outcomes should be conditional on all variables that affect the probability of inclusion, which are the variables used in survey weighting. Regression models suffer from computational problems with deep interaction. We would like to incorporate these weighting variables into the model for survey outcomes under the framework of multilevel regression and poststratification at much finer levels. Our procedure will yield the model-based weights after smoothing, which are evaluated via simulations comparing with classical weights. We use Stan for computation and illustrate the performances via the application of the New York Longitudinal Survey of Poverty study.

e-mail: y*si@biostat.wisc.edu*

ROBUST BAYESIAN MODELS FOR SURVEYS WITH MISSING DATA AND EXTERNAL INFORMATION

Sahar Z. Zangeneh*, Fred Hutchinson Cancer Research Center

Roderick J.A. Little, University of Michigan

Survey data are often collected according to some complex probability sampling design. Design-based inference aims at obtaining unbiased, or minimally biased, weighted estimates of finite population quantities such as means or totals, with respect to the sampling design. Within this framework, calibration methods are often employed to incorporate external information. These methods re-adjust the original weights to satisfy constraints imposed by the external information while deviating minimally from the original weights. However, such estimators are sensitive to the choice of distance measure. An alternative paradigm is Bayesian model-based estimation, where a model is chosen to predict the non-observed data. Such models need to incorporate features of the sampling design, like sampling weights and clustering, to be robust to model misspecification. We describe two applications of this approach to

survey problems involving external information: (i) probability proportional to size sampling, where aggregate information is available for sizes of non-sampled units, and (ii) survey nonresponse, when there is external information for post-stratification. Simulation comparisons with standard "design-based" methods suggest superior frequentist properties for the Bayesian method.

e-mail: *saharzz@fhcrc.org*

125. CAUSAL INFERENCE IN SOCIAL NETWORKS

CAUSAL ESTIMATION OF PEER EFFECTS IN LONGITUDINAL DYADIC DATA USING INSTRUMENTAL VARIABLES

A. James O'Malley*, Geisel School of Medicine at Dartmouth

Felix Elwert, University of Wisconsin, Madison

J. Niels Rosenquist, Massachusetts General Hospital

Alan M. Zaslavsky, Harvard Medical School

Nicholas A. Christakis, Yale University

The identification of causal peer effects (social contagion or induction) from observational data in social networks is challenged by two distinct sources of bias: latent homophily and unobserved confounding. Directed acyclic graphs (DAGs) of data generating models encompassing both homophily and confounding are used to investigate whether peer effects of behaviors can be identified using genes as instrumental variables (IVs). Three main identification results obtained. First, using a single fixed gene as an IV will generally fail to identify peer effects if the gene affects past values of the treatment. Second, multiple fixed genes, or, more promisingly, time-varying gene expression can identify peer effects if we instrument exclusion violations as well as the focal treatment. Third, we show that IV identification of peer effects remains possible even under multiple complications often regarded as lethal for IV identification of intra-individual effects, such as pleiotropy on observables, homophily on past phenotype, past and ongoing homophily on genotype, inter-phenotype peer effects, population stratification, gene expression that is endogenous to past phenotype and past gene expression, and others. We apply the methodology to estimate peer effects of body mass index (BMI) among friends and spouses in the Framingham Heart Study.

e-mail: *Alistair.J.O'Malley@dartmouth.edu*

OBSERVATIONAL CAUSAL INFERENCE IN COMMUNITY-STRUCTURED SOCIAL NETWORKS

Cosma R. Shalizi*, Carnegie Mellon University

Contagion or social-influence effects are generally unidentified in observational studies of social networks, because of the presence of latent homophily. On the other hand, under many plausible models, networks which form through homophily should divide into "modules" or "communities" of densely-connected nodes, which can be recovered through clustering. I will report some theoretical results on when controlling on estimated community membership will (asymptotically) yield unconfounded estimates of contagion effects, and some simulation evidence about how practical the idea is.

e-mail: *cshalizi@cmu.edu*

INDIRECT ADJUSTMENT FOR HOMOPHILY BIAS WITH A NEGATIVE CONTROL VARIABLE IN PEER EFFECT ANALYSIS

Lan Liu*, Harvard University

Eric Tchetgen Tchetgen, Harvard University

Analysis with social network data have suggested that health related outcomes such as obesity and psychological states such as happiness and loneliness can be spread over a network. However, such analysis of peer effects or contagion effects have come under critique. The challenge in such analysis lies in adjusting for homophily bias for the sample collected. Negative control variables are useful in adjusting bias in study where individuals are not selected at random. Negative control treatment is defined to be an exposure variable that is known not have a direct effect on the outcome of interest and negative control outcome is defined to be a response variable that is known not to be directed affected by the treatment of interest. In this paper, we propose indirect adjustment methods with negative control variables to estimate the peer effect in study with homophily bias. A simulation study is carried out to investigate the finite sample performance of the proposed estimators. The methods are further illustrated in an application.

e-mail: *lanl@email.unc.edu*

SEGREGATED GRAPHS AND MARGINALS OF CHAIN GRAPH MODELS

Ilya Shpitser*, Johns Hopkins University

Bayesian networks are a popular representation of asymmetric (for example causal) relationships between random variables. Markov random fields (MRFs) are a complementary model of symmetric relationships used in computer vision, spatial modeling, and social and gene expression networks. A chain graph model under the Lauritzen-Wermuth-Frydenberg interpretation generalizes both Bayesian networks and MRFs, and can represent asymmetric and symmetric relationships together. We show that special mixed graphs which we call segregated graphs can be associated, via a Markov property, with supermodels of a marginal of chain graphs

defined only by conditional independences. Special features of segregated graphs imply the existence of a very natural factorization for these supermodels, and imply many existing results on the chain graph model, and the ordinary Markov model carry over, including parameter fitting procedures. We show, via a simulation study, how interference between unit outcomes can be directly encoded as parameters in a segregated graph model.

e-mail: *ilyas@cs.jhu.edu*

126. OPTIMAL DESIGN FOR NONLINEAR MODELS

A BAYESIAN DECISION THEORETIC APPROACH TO EXPERIMENTAL DESIGNS FOR HORMESIS

Steven B. Kim*, California State University, Monterey Bay

Scott M. Bartell, University of California, Irvine

Daniel L. Gillen, University of California, Irvine

Hormesis is a non-monotonic dose-response relationship which can be characterized by a beneficial effect at low dose and a harmful effect at high dose. Though scientists have been interested in hormesis, many existing data suffer from poor experimental designs in order to test for hormesis. High sensitivity for detecting hormesis requires a large sample size and a large number of experimental doses together with a careful experimental design. In this manuscript we consider the application of Bayesian decision theory to allocate experimental units for hypothesis testing for hormesis. We consider two loss functions, one devised from precision estimation of a parameter of interest and the other one devised from information gain at low doses, particularly in an assumed hormetic zone. While the former method requires a parametric dose-response model, the latter method does not require a parametric assumption. In simulation studies, we show that the Bayesian decision theoretic approaches outperform the typical experimental designs in the past. We further discuss the effect of the number of experimental doses and the length of a hormetic zone relative to an experimental range.

email: *stkim@csumb.edu*

OPTIMAL DESIGN FOR DOSE-FINDING STUDY WITH DELAYED RESPONSES

Tian Tian, University of Illinois, Chicago

Lei Nie, U.S. Food and Drug Administration

Min Yang*, University of Illinois, Chicago

Delayed response is not uncommon in Phase I clinical trials. We may not be able to observe the toxicity outcome before the next dose assignment. Waiting for each patient's outcome may result in unfeasibly long trial. Ignoring the unobserved data may underestimate the toxicity and would put patients in danger of overly toxic doses. A proper approach is to treat the unobserved responses as censored observations. While there is substantial optimality literature for Phase I clinical trials, those results mainly focus on the situation that there is no delayed response. Little is known how to choose an optimal/efficient design when delayed responses are presented. In this talk, we shall systematically develop a new framework to address this issue. Some efficient/robust designs will be derived under this framework.

email: *minyang.stat@gmail.com*

NATURE-INSPIRED META-HEURISTIC ALGORITHMS FOR GENERATING OPTIMAL DESIGNS FOR NONLINEAR MODELS

Weng Kee Wong*, University of California, Los Angeles

Nature-inspired meta-heuristic algorithms are increasingly studied and used in many disciplines to solve high-dimensional complex optimization problems in the real world. It appears relatively few of these algorithms are used in mainstream statistics even though they are simple to implement, very flexible and able to find an optimal or a nearly optimal solution quickly. Frequently, these methods do not require any assumption on the function to be optimized and the user only needs to input a few tuning parameters. I will demonstrate the usefulness of some of these algorithms for finding different types of optimal designs for nonlinear models in dose response studies. Algorithms that I plan to discuss are more recent ones such as Cuckoo and Particle Swarm Optimization. I also compare their performances and advantages relative to deterministic state-of-the art algorithms.

email: *wkwong@ucla.edu*

127. BAYESIAN METHODS

BAYESIAN REGRESSION ANALYSIS FOR ESTIMATING DISEASE ETIOLOGY

Zhenke Wu*, Johns Hopkins Bloomberg School of Public Health

Scott L. Zeger, Johns Hopkins Bloomberg School of Public Health

In epidemiology studies of possible causes of disease, multivariate binary biomarker or clinical data are routinely collected on both cases and controls. One example is the Pneumonia Etiology Research for Child Health (PERCH) study where the presence/absence of more than 30 pneumonia-causing pathogens are simultaneously measured in blood, sputum and the nasal cavity with an objective to estimate the fraction of cases caused by each pathogen, termed "etiologic fractions". Wu, Deloria-Knoll, Hammitt and Zeger (2015, JRSS-C) developed and applied latent variable models to estimate etiology from distributional differences between cases' and controls'

measurements. An important scientific goal is to describe the effects of explanatory variables on disease etiology. In addition, adjustment for covariates may be necessary if there exists differences between cases' and controls' covariate distributions. In this paper, we propose a family of latent variable regression models to address these aims. The models represent the distribution of each case's observation as a mixture of components, each one representing a single cause. We let the mixing weights vary with continuous and discrete covariates. We also allow the false positive error rates to vary with covariates in a separate regression. The mixture component distributions are partly informed by control samples. We derive the model properties and then apply the model to the motivating PERCH study to characterize the dependence of the causes of pneumonia on season, age, HIV status and other variables. All the simulations and analyses are enabled by a new software package "baker" at https://github.com/zhenkewu/baker.

email: *zhwu@jhu.edu*

CONTROLLING FOR SYSTEMATIC BIAS IN ALLELIC IMBALANCE ESTIMATION USING A NEGATIVE BINOMIAL BAYESIAN MODEL

Luis G. Leon Novelo*, University of Texas School of Public Health
Lauren M. McIntyre, University of Florida College of Medicine
Alison R. Gerken, University of Florida College of Medicine
Alison M. Morse, University of Florida College of Medicine

Justin M. Fear, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health
Sergey Nuzhdin, University of Southern California

A gene is in allelic imbalance (AI) when its two alleles have different expression levels. We propose a Bayesian model to detect genes in AI using RNA next generation sequencing data. The experiment consists in crossing two near isogenic lines of Drosophila melanogaster, a naturally derived genotype from female (line) to W1118 male flies (tester). The F1 heterozygous female offspring are separated by sex immediately after enclosure and half of the flies are kept virgin while the other half are mated to W1118. Per gene, the number of reads aligning to the tester allele, line allele and both alleles is recorded. The gene is in AI if the proportion of reads produced by the tester is different from the one produced by the line. The challenges are: (i) the sample proportion of reads aligning to the tester is not necessary an unbiased estimate of the proportion of reads produced by the tester due to potential systematic bias, (ii) the coverage can be different between mated and virgin environments, and (iii) data are overdispersed. We address (i) using simulation to estimate the degree of systematic bias, (ii) including information of reads aligning to both alleles, and (iii) using a negative-binomial sampling distribution.

email: *Luis.G.LeonNovelo@uth.tmc.edu*

PATIENT-SPECIFIC PREDICTION OF ABDOMINAL AORTIC ANEURYSM EXPANSION USING BAYESIAN CALIBRATION

Liang Liang Zhang*, Michigan State University
Justin Mrkva, Michigan State University
Sajjad Seyedsalehi, Michigan State University
Jongeun Choi, Michigan State University
Chae Young Lim, Seoul National University
Tapabrata Maiti, Michigan State University
Seungik Baek, Michigan State University

Translating recent advances in abdominal aortic aneurysm (AAA) growth and remodeling (G&R) into a predictive, patient-specific clinical treatment tool requires a major paradigm shift in computational science. The aims of this paper are to develop a prediction framework that 1) first calibrates the physical AAA G&R model using patient-specific longitudinal computed tomography (CT) scan images, 2) predicts the expansion of an AAA in future time, and 3) analyzes the associated uncertainty in the prediction. To achieve our aims, we first formulate the Bayesian calibration of our AAA G&R computation model taking into account model inadequacy, prior distributions of model parameters, measurement errors, and patient-specific longitudinal CT scan images. Next, we demonstrate how to achieve the proposed aims by solving the formulated Bayesian calibration problems for cases with the synthetic G&R model output data and real medical patient-specific CT data. In particular, we compare and discuss the performance of predictions and computation time under different sampling cases of the model output data and patient data, both of which are simulated by the G&R computation. Finally, we apply our Bayesian calibration to real patient-specific longitudinal CT data and validate our prediction, showing the usefulness of our approach to the computational science and medical communities.

email: *lions_z@hotmail.com*

SPATIAL SKEW-NORMAL/INDEPENDENT MODELS FOR CLUSTERED PERIODONTAL DATA WITH NON-RANDOM MISSINGNESS

Dipankar Bandyopadhyay*, Virginia Commonwealth University
Victor H. Lachos, University of Campinas, Brazil
Marcos Prates, Universidade Federal de Minas Gerais, Brazil
Xioayue Zhao, University of Minnesota

Periodontal studies often leads to data collected which are clustered in nature viz. clinical attachment level (or CAL), that are routinely analyzed under a linear mixed model framework with underlying normality assumptions of the random effects and random errors. However, a careful look reveals that these data might exhibit skewness and tail behavior, and hence the usual normality

assumptions might be questionable. In addition, periodontal progression might also be spatially associated, i.e. proximal tooth sites seem to have similar disease status that sites located further away. Furthermore, the presence/absence of a tooth is informative as the number and location of missing teeth informs about the periodontal health in that region. In this talk, we develop a (shared) random effects model for site-level CAL and binary presence/absence status of a tooth under a Bayesian paradigm. The random effects are modelled using a spatial skew-normal/independent (S-SNI) distribution, whose covariance structure follows a conditionally autoregressive (CAR) density. Our S-SNI density provides an attractive parametric alternative to model spatially referenced asymmetric thick-tailed structures. Both simulation studies and analysis of a real dataset reveal significantly improved fits over models that do not consider these features.

email: *dbandyop@vcu.edu*

BAYESIAN APPROACH FOR CLUSTERED INTERVAL-CENSORED DATA WITH TIME-VARYING COVARIATE EFFECTS

Yue Zhang*, University of Cincinnati

Xia Wang, University of Cincinnati

Bin Zhang, Cincinnati Children's Hospital Medical Center

Interval censored data arise when failure times cannot be observed exactly but can only be determined to lie within an interval. Interval censored data are very common in clinical trials and epidemiological studies. This paper considered a Bayesian approach for correlated interval censored data under a dynamic Cox regression model when random effect presents. Some methods have been developed for clustered data with temporal covariate effects. However, they are limited to right censoring. In this paper, we estimated piecewise constant coefficients based on a dynamic Cox regression model under Bayesian framework. The dimensions of coefficients were automatically determined by reversible jump Markov chain Monte Carlo algorithm. Meanwhile, we used a shared frailty factor for unobserved heterogeneity or for statistical dependence between the observations. Simulation studies with various combinations of time-varying coefficients and frailties were conducted to verify our method. We also applied this methodology to a children behavior development study to demonstrate the properties of the method.

email: *zhang3ye@mail.uc.edu*

REPULSIVE PRIORS FOR MEANINGFUL INFERENCES IN BIOMEDICAL APPLICATIONS

Yanxun Xu*, Johns Hopkins University

Peter Mueller, University of Texas, Austin

Donatello Telesca, University of California, Los Angeles

We discuss the use of the determinantal point process (DPP) as a prior for latent structure in biomedical applications, where inference often centers on the interpretation of latent features as biologically or clinically meaningful structure. Typical examples include mixture models, when the terms of the mixture are meant to represent clinically meaningful subpopulations (of patients, genes, etc.). Another class of examples are feature allocation models. We propose the DPP prior as a repulsive prior on latent mixture components in the first example, and as prior on feature-specific parameters in the second case. We argue that the DPP is in general an attractive prior model for latent structure when biologically relevant interpretation of such structure is desired. We illustrate the advantages of DPP prior in three case studies, including inference in mixture models for magnetic resonance images (MRI) and for protein expression, and a feature allocation model for gene expression using data from The Cancer Genome Atlas. An important part of our argument are efficient and straightforward posterior simulation methods. We implement a variation of reversible jump Markov chain Monte Carlo simulation for inference under the DPP prior, using a density with respect to the unit rate Poisson process.

email: *yxu.stat@gmail.com*

A MODEL AND R PACKAGE FOR BAYESIAN SURVIVAL AND MULTISTATE ANALYSIS

Adam King*, California State Polytechnic University, Pomona

We describe a framework for analyzing general event time or event history data, which includes survival and multistate problems as special cases. We model cause-specific hazards for multiple event types and multiple at-risk states as functions of time-varying fixed and random effects, with Gaussian Markov random field priors used for semiparametric smoothing of these effects. Inferences are obtained using a block Metropolis-Hastings routine that requires no user tuning. We illustrate an R implementation of this model using multistate data on recurrent illicit drug use.

email: *king@cpp.edu*

128. CAUSAL INFERENCE IN EPIDEMIOLOGY AND HEALTH POLICY

DEFINING AND ESTIMATING CAUSAL DIRECT AND INDIRECT EFFECTS WHEN SETTING THE MEDIATOR TO SPECIFIC VALUES IS NOT FEASIBLE

Judith J. Lok*, Harvard School of Public Health

Natural direct and indirect effects decompose the effect of a treatment into the part that is mediated by a covariate (the mediator)

and the part that is not. Their definitions rely on the concept of outcomes under treatment with the mediator "set" to its value without treatment. Typically, the mechanism through which the mediator is set to this value is left unspecified, and in many applications it may be challenging to fix the mediator to particular values for each unit or individual. Moreover, how one sets the mediator may affect the distribution of the outcome. This article introduces "organic" direct and indirect effects, which can be defined and estimated without relying on setting the mediator to specific values. Organic direct and indirect effects can be applied for example to estimate how much of the effect of some treatments for HIV/AIDS on mother-to-child transmission of HIV-infection is mediated by the effect of the treatment on the HIV viral load in the blood of the mother.

email: *jlok@hsph.harvard.edu*

USING STRUCTURAL-NESTED MODELS TO ESTIMATE THE EFFECT OF CLUSTER-LEVEL ADHERENCE ON INDIVIDUAL-LEVEL OUTCOMES WITH A THREE-ARMED CLUSTER-RANDOMIZED TRIAL

Babette A. Brumback*, University of Florida

Zhulin He, Iowa State University

Shanjun Helian, University of Florida

Matthew Freeman, Emory University

Richard Rheingans, University of Florida

Much attention has been paid to estimating the causal effect of adherence to a randomized protocol using instrumental variables to adjust for unmeasured confounding. Our interest stems from a wish to estimate the effect of cluster-level adherence on individual-level binary outcomes with a three-armed cluster-randomized trial and polytomous adherence. We recently developed two structural-nested modeling approaches for estimation; the approaches differ in the handling of measured individual-level confounders of the effect of randomization on the outcome. The first approach uses a weighted generalized structural nested mean model, which adjusts for the confounders using weights, and the second approach uses an ordinary generalized structural nested mean model, which stratifies on the confounders. The two approaches target different estimands. Our methodology accommodates cluster-randomized trials with unequal probability of selecting individuals. Furthermore, we developed a method to implement the approaches with relatively simple programming. The approaches work reasonably well, but when the structural-nested model does not fit the data, there may be no solution to the estimating equation. We investigate the performance of the approaches using simulated data, and we also use them to estimate the effect on pupil absence of school-level adherence to a randomized water, sanitation, and hygiene intervention in western Kenya.

email: *brumback@ufl.edu*

A MULTIPLE-IMPUTATION BASED DOUBLY ROBUST ESTIMATION OF TREATMENT EFFECTS IN LONGITUDINAL STUDIES

Tingting Zhou*, University of Michigan

Michael Elliott, University of Michigan

Roderick Little, University of Michigan

Because observational studies usually lack randomization, valid inference about causal effects can only be drawn by controlling for confounders. However, when time dependent confounders are present, adjusting for such confounders in a traditional regression model can be inadequate, since such confounders can serve as mediators of treatment effects. In this study, we develop a Bayesian approach to causal inference, called Penalized Spline of Propensity Prediction (PSPP) (Zhang & Little 2008). PSPP was originally proposed for missing data problems. Here, we extend the method to estimate causal effects, by imputing missing potential outcomes and drawing inference based on imputed and observed outcomes. PSPP relies on the balancing property of propensity score to achieve double robustness by modelling the relationship between propensity scores and outcomes as a penalized spline regression. In simulation studies, we demonstrate that PSPP yields consistent estimates of causal effects. PSPP tends to perform better than doubly robust MSM models (Bang & Robins 2005, Bryan & van der Laan 2004) when the relationship between propensity score and outcome is nonlinear or when the weights are highly variable. Our method is also used to evaluate the effects of time dependent interventions for low urine output due to kidney injury on survival.

email: *tkzhou@umich.edu*

A GENERAL APPROACH ON CAUSAL MEDIATION ANALYSIS

Pan Wu*, Christiana Care Health System

In biomedical, social science, and healthcare studies, researchers usually want to know if an intervention or risk factor exposure has a causal direct effect on outcomes that is mediated by certain post-treatment factors. Since early 1980s, a number of different approaches and assumptions on mediation analysis have been presented with an emphasis on identification and estimation of medication effects, such as Linear Structural Equations Models (LSEM), Marginal Structural Models (MSM), and some confounding assumptions including sequential ignorability assumption and no-interaction assumption between treatment and mediator. In this talk, I will introduce a new approach in estimation of causal mediation effects under the functional response models and the potential outcome framework with no unmeasured confounding and sequential ignorability assumptions. The new approach has robust inference, good asymptotic properties, and light computational workload. It can easily be extended to complex experiment

designs and non-continuous mediator and outcome within non-linear models as well.

email: *pan.wu@outlook.com*

IMPROVING COVARIATE BALANCING PROPENSITY SCORE FOR CONTINUOUS TREATMENT REGIMES

Samantha Noreen*, Emory University

Qi Long, Emory University

The propensity score plays an essential role in causal inference using observational data. However, a number of challenges arise when using the propensity score to deal with non-binary treatment regimes. In this work, we propose an approach to improve the covariate balancing propensity score (CBPS; Imai and Ratkovic, 2014; Fong et al. 2015) for continuous treatment regimes that can handle both continuous and discrete covariates. The proposed approach is shown to outperform the original CBPS in simulations and is further illustrated through analysis of the clinical data from the Emory Amyotrophic Lateral Sclerosis (ALS) Center.

email: *snoreen@emory.edu*

CALIBRATE MEASUREMENT ERRORS AND MISCLASSIFICATIONS IN MENDELIAN RANDOMIZATION STUDIES

Cheng Zheng*, University of Wisconsin, Milwaukee

Mendelian randomization is a powerful tool to study the causal relationship between a phenotype and a health outcome when some of the confounding factors are unmeasured. In practice, the collected phenotype data, such as dietary pattern and physical activity, are often from survey data and face the problem of measurement error or misclassification that with systematic bias related to personal characteristics. For some applications, it is impossible to obtain the gold standard value for exposure. What we can do is to have a relatively small subset of the cohort with objectively measured exposure that contains only non-differential measurement error. Moreover, the genotype or the genetic risk scores could also contain variations. Using genotype as instrumental variable, two commonly used estimators for the causal effect of phenotype on a continuous health outcome are two stage least squares (2SLS) and generalized method of moment (GMM). We propose regression calibration estimator to handle the measurement error and misclassification in the phenotype and genotype for both two stages of regressions. We also propose an estimator from the corrected GMM estimating equations to handle measurement error issue. We derived the asymptotic for our proposed estimators and the simulation studies suggested that our proposed estimator has good performance with finite sample.

email: *zhengc@uwm.edu*

129. COUNT AND CATEGORICAL DATA ANALYSIS

A NEW COMPOUND CLASS OF EXPONENTIATED POWER LINDLEY-LOGARITHMIC DISTRIBUTION: MODEL, PROPERTIES AND APPLICATIONS

Mavis Pararai*, Indiana University of Pennsylvania

Jacinth A. Maynard, Lock Haven University of Pennsylvania

Gayan W. Liyanage, Central Michigan University

A new class of distributions called the Exponentiated Power Lindley-Logarithmic (EPLL) distribution is introduced and its properties are explored. This new distribution represents a more flexible model for lifetime data. Some statistical properties of the proposed distribution including the expansion of the density function, quantile function, hazard and reverse hazard functions, moments, conditional moments, moment generating function, skewness and kurtosis are presented. Mean deviations, Bonferroni and Lorenz curves, Renyi entropy and distribution of the order statistics are derived. Maximum likelihood estimation technique is used to estimate the model parameters. A simulation study is conducted to examine the bias, mean square error of the maximum likelihood estimators and width of the confidence interval for each parameter and nally applications of the model to real data sets are presented to illustrate the usefulness of the proposed distribution.

email: *pararaim@iup.edu*

A BAYESIAN APPROACH IN ESTIMATING ODDS RATIOS FOR RARE OR ZERO EVENTS

Mehmet Kocak*, University of Tennessee Health Science Center

For rare events, one or two cells of 2x2 tables may have zero (0) counts or a count very close to zero. This issue makes the estimation of Odds Ratio (OR) impossible or leads to unstable OR estimates due to complete or quasi-complete separation. In such cases, Exact approaches or Firth approach may be helpful but these two methods do not remove the issue completely. In this research, we propose a Bayesian estimation procedure to estimate OR and provide a test procedure that is much more powerful than the Exact Method and Firth approach at the same Type-1 Error level.

email: *mkocak1@uthsc.edu*

ANALYSIS OF INFLATED BIVARIATE COUNT DATA THAT OCCUR IN HEALTH CARE STUDIES USING POISSON REGRESSION MODELS

N. Rao Chaganty*, Old Dominion University

Pooja Sengupta, International Management Institute

Bivariate count data are common in health care studies such as twin or cross over studies. For instance, Carlin et al. (1987, Journal of Pediatrics) present bivariate data that consists of the number of infections in both ears of toddlers over a period of six months. A careful examination of the data shows that the frequencies of (0,0) and (2,2) cells are high. The inflation at (0,0) was due to a large number of toddlers who were never infected in either ear. Interestingly, pediatricians noted that (i) infections in one ear often spread to the other ear, and (ii) infected toddlers had a high probability of a relapse infection at a later date. Thus, there was also an increased frequency of (2,2) in the data. Another example is data from the Australian health survey of 1977-78 (Cameron et al. 1988). The data consists of bivariate count responses of the number of doctor visits and the number of medicines prescribed. Here there was again an inflated frequency of (0,0), as many patients did not visit their physician and thus received no prescriptions, but there was also an inflated frequency of (1,1), where patients saw their physician once and had one prescription. To analyze these data we introduce bivariate doubly-inflated Poisson regression models. We discuss distributional properties, parameter estimation and goodness of fit.

email: *rchagant@odu.edu*

A BAYESIAN TEST OF INDEPENDENCE IN A TWO-WAY CONTINGENCY TABLE WITH COVARIATES UNDER CLUSTER SAMPLING

Dilli Bhatta*, University of South Carolina Upstate

Balgobin Nandram, Worcester Polytechnic Institute

We consider a Bayesian approach for the test of independence to study the association between two categorical variables from a two-stage cluster sampling design. We incorporate the covariates at both unit and cluster levels in the test. Our main idea for the Bayesian test of independence is to convert the cluster sample with covariates into an equivalent simple random sample without covariates which provides a surrogate of the original sample. Then, this surrogate sample is used to compute the Bayes factor to make an inference about independence. We apply our methodology to the data from the Trend in International Mathematics and Science Study (2007) for fourth grade U.S. students to assess the association between the mathematics and science scores represented as categorical variables and also provide the simulation study. The result shows that if there is strong association between two categorical variables, there is no significant difference between the tests with and without the covariates. However, in the simulation study, we found noticeable difference in borderline cases (moderate association between the two categorical variables).

email: *dbhatta@uscupstate.edu*

SIMULATING LONGER VECTORS OF CORRELATED BINARY RANDOM VARIABLES VIA MULTINOMIAL SAMPLING

Justine Shults*, University of Pennsylvania Perelman School of Medicine

The ability to simulate correlated binary data is important for sample size calculation and comparison of methods for analysis of clustered and longitudinal data with dichotomous outcomes. Sampling from the multinomial distribution of all possible length n permutations of zeros and ones is a straightforward simulation approach that was first proposed by Kang and Jung (Biom. J. 2001; 43 (3): 1521-4036). However, the multinomial sampling method has only been implemented in general form (without first making restrictive assumptions) for vectors of length 2 and 3. As noted by Haynes, Sabo, and Chaganty (2015) "the CDF for establishing decision rules becomes complicated for cases of four or more repeated measures. While not impossible, constructing higher order joint probabilities can be computationally challenging." In this presentation I present an algorithm for simulating correlated binary data via multinomial sampling that can be applied to directly compute the joint distribution for any n. I demonstrate my algorithm to simulate vectors of length 4 and 8 in an assessment of power during the planning phases of a study and to assess the choice of working correlation structure in an analysis with GEE.

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

TESTING FOR TREND WITH A NOMINAL OUTCOME

Aniko Szabo*, Medical College of Wisconsin

The Cochran-Armitage test is commonly used to test for trend with binary outcomes, however there is no established procedure for a trend test with nominal outcomes that would provide both a global hypothesis test and outcome-specific inference. We derive a simple formula for such a test using a weighted sum of Cochran-Armitage test statistics evaluating the trend in each outcome separately. The test is shown to be equivalent to the score test for multinomial logistic regression. The new formulation enables the derivation of explicit sample size formulas and multiplicity-adjusted inference for individual outcomes. Extensions to clustered and survey-weighted data are discussed.

email: *aszabo@mcw.edu*

ESTIMATION OF THE OPTIMAL ROC IN COMPLEX CLASSIFICATION SETTINGS

WITHDRAWN

Daniel B. Shin*, University of Pennsylvania

Farrah J. Mateen, Massachusetts General Hospital and Harvard Medical School

Jaroslaw Hareslak, Indiana University, Indianapolis

Joel M. Gelfand, University of Pennsylvania
Russell T. Shinohara, University of Pennsylvania

Simple classification scenarios involving thresholding a predictor using a cutpoint offer straightforward assessments of classifier performance using the receiver operating characteristic (ROC) curve, but for complex scenarios, such as bivariate thresholding of a predictor using upper and lower cutpoints, a classifier is inadequately defined and the ROC curve not specified. We introduce a more rigorous definition of a classifier and propose a generalization of the ROC analysis called the optimal ROC curve (OROC), which has similar properties to the standard ROC curve. The nonparametric estimation procedure of OROC simultaneously estimates the newly defined OROC classifier. Alternative semiparametric and parametric methods for the OROC classifier estimation are presented: the generalized additive model (GAM), and the maximum likelihood estimation (MLE) based on a profile likelihood. In Monte Carlo simulations, the OROC and GAM classifiers consistently performed better than the MLE classifier. In our motivating example of serum sodium levels in patients hospitalized for fulminant bacterial meningitis, the three classifiers performed similarly using complete data, but cross-validation using subsampling at random showed differential performance at small sample sizes. Complex classification can be generalized beyond bivariate thresholding to include multiple thresholds, and augment existing classification scenarios involving molecular and imaging biomarkers.

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

130. JOINT MODELS FOR LONGITUDINAL AND SURVIVAL DATA

WEIGHTED ZIP MIXED MODEL WITH AN APPLICATION TO MEDICAID DATA

Sang Mee Lee*, University of Chicago
Theodore Karrison, University of Chicago

In medical or biological research, it is common to encounter clustered count data with excess zeros. For example, health care utilization data are often found multi-modal with excess zeroes as well as multilevel structure where patients are nested within physicians and hospitals. Zero-inflated count models with random effects have been developed to analyze this type of data. However, no study has considered a situation where data are censored due to the finite nature of the observation period or follow-up. In this paper, we present a weighted version of zero-inflated poisson model with random effects accounting for variable individual follow-up times. The performance of the proposed model is evaluated through simulation studies and we apply our approach to Medicaid data analysis.

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

A SEMIPARAMETRIC JOINT MODEL FOR LONGITUDINAL DATA AND SURVIVAL IN END-OF-LIFE STUDIES

Zhigang Li*, Dartmouth College
H. R. Frost, Dartmouth College
Lihui Zhao, Northwestern University
Lei Liu, Northwestern University
Kathleen D. Lyons, Dartmouth College
Huaihou Chen, University of Florida
Bernard Cole, University of Vermont
David Currow, Flinders University, Australia
Marie Bakitas, University of Alabama
Tor D. Tosteson, Dartmouth College

In end-of-life medical research studies, it is often of interest to study quality of life (QOL) of patients during a relatively short period (e.g., 1 year) prior to death. Survival duration is commonly seen as a secondary outcome. Longitudinal measurements of QOL are commonly collected until death or study end and life expectancy is relatively short in such studies. In analysis of these data, the interplay between longitudinal and survival outcomes should be taken into account since these two outcomes are associated with each other especially during end-of-life period. Without appropriate handling of the association, clinical meaning of the estimates could be distorted and inefficient and even biased results could be generated. Censoring of the survival times occurs frequently and makes it challenging to account for the relationship between the longitudinal outcomes and survival time. To address these issues, we propose a novel semiparametric approach using natural cubic splines for jointly modeling longitudinal and survival data in end-of-life medical research studies. A semiparametric mixed effects submodel for the longitudinal data and a time-varying coefficient Cox submodel for the survival data are used. Simulation study is also carried out to explore the finite sample properties of the estimators.

email: *zhigang.li@dartmouth.edu*

DYNAMIC PREDICTION FOR MULTIPLE REPEATED MEASURES AND EVENT TIME DATA: AN APPLICATION TO PARKINSON'S DISEASE

Jue Wang*, University of Texas Health Science Center, Houston
Sheng Luo, University of Texas Health Science Center, Houston
Liang Li, University of Texas MD Anderson Cancer Center

In many clinical trials studying neurodegenerative diseases such as Parkinson's disease (PD), multiple longitudinal outcomes are collected to fully explore the multidimensional impairment caused by this disease. If the outcomes deteriorate rapidly, patients may reach a level of functional disability sufficient to initiate levodopa

therapy for ameliorating disease symptoms. An accurate prediction of the time to functional disability is helpful for physicians to monitor patients' disease progression and make informative medical decisions. In this article, we first propose a joint model that consists of a latent trait linear mixed model (LTLMM) for the multiple longitudinal outcomes, and a survival model for event time. The two submodels are linked together by an underlying latent variable. We develop a Bayesian approach for parameter estimation and a dynamic prediction framework for predicting target patients' future outcome trajectories and risk of a survival event, based on their multivariate longitudinal measurements. Our proposed model is evaluated by simulation studies and is applied to the DATATOP study, a motivating clinical trial assessing the effect of deprenyl among patients with early PD.

email: *Jue.Wang@uth.tmc.edu*

FLEXIBLE LINK FUNCTIONS IN A JOINT MODEL OF BINARY AND LONGITUDINAL DATA

Dan Li*, University of Cincinnati

Xia Wang, University of Cincinnati

Seongho Song, University of Cincinnati

Nanhua Zhang, Cincinnati Children's Hospital Medical Center

Dipak K. Dey, University of Connecticut

Joint models for the association of a binary primary endpoint and a longitudinal continuous process are proposed when their association is of interest. The dependence between these two processes can be characterized by introducing a common set of latent random effects. Due to the binomial nature of the primary endpoint, an important consideration that has been less investigated is to choose appropriate link functions for the joint model. We introduce two families of flexible link functions based on the generalized extreme value (GEV) distribution and the symmetric power logit (splogit) distribution. Our work is the first to investigate the importance of an appropriate and flexible link function in improving the estimation and prediction of a Bayesian joint model. The parameters are estimated using Markov chain Monte Carlo (MCMC) methods. Flexibility and gains of the proposed joint model are demonstrated through detailed studies on simulated data sets and an application to a real data example about how long-term marital stress affects late-life major depressive disorder in a group of aging women.

email: *lid7@mail.uc.edu*

JOINT MODELING OF FUNCTIONAL DATA AND TIME TO EVENT: AN APPLICATION TO FECUNDITY STUDIES

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

Animikh Biswas, University of Maryland Baltimore County

In longitudinal studies, it is often of interest to investigate how the functional features of a marker's measurement process is associated with the event time of interest. We make use of B-splines to smoothly approximate the infinite dimensional functional data and propose a joint model of the longitudinal functional features and the time to event. The proposed approach also allows for prediction of survival probabilities for future subjects based on their available longitudinal measurements and a fitted joint model. We illustrate our proposals on a prospective pregnancy study, namely Oxford Conception Study, where hormonal measurements of luteinizing hormone, estrogen indicate timing of ovulation and whether ovulation occurs in a menstrual cycle. A joint modeling approach using functional analytic approach and discrete survival modeling was used to assess whether the functional features of hormonal measurements, such as the peak of the hormonal profile, its curvatures as well as timing of peak are associated with time to pregnancy. This is based on joint work with Dr. Rajeshwari Sundaram and Dr.Animikh Biswas.

email: *mlbegood@gmail.com*

A JOINT MODEL APPROACH FOR LONGITUDINAL DATA WITH NO TIME ZERO AND TIME-TO-EVENT WITH A COMPETING RISK

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

Station is a digitalized measure of how low the fetus' head is positioned in the pelvis of a pregnant woman. It is measured from -3 to +4, where a value of -3 implies a fetus is still very high in the pelvis and not close to delivery, while +4 implies the fetus is below the pelvis and is due for delivery. The fetus is delivered vaginally (spontaneous or vacuum) or through a C-section. It is of interest to predict the timing and delivery type using individualized longitudinal assessments of station. Importantly, women enter the hospital at different station measurements, resulting in no clear time zero for use as a reference point for valid statistical inferences and predictions. We develop a shared random parameter model that links together model components for the longitudinal station process (with no time-zero) and both the time and type of delivery. Specifically, each model component includes random effects that are shared between these three components, inducing realistic dependence between these three data elements. The goal in constructing this model is to develop an adaptive predictor to predict both the timing and type of delivery based on repeated station values. We construct

a Bayesian approach for parameter estimation that can be implemented with STAN code. The approach is illustrated using a longitudinal cohort of digitized station measurements and the timing and type of delivery in an international cohort. We evaluate the prospect performance of our approach using cross-validation.

email: *olive.buhule@nih.gov*

131. PERSONALIZED MEDICINE

COMPARING MOBILE HEALTH TREATMENT POLICIES

Peng Liao*, University of Michigan

Pedja Klasjna, University of Michigan

Susan A. Murphy, University of Michigan

Mobile devices can be used to provide treatment or support whenever needed and adapted to the context of the user. The treatment policies, also known as Just-In-Time Adaptive Interventions (JITAI), are composed of decision rules that, at each decision time, map user's context (e.g. location, weather, current time, social activity, stress, and urges to smoke) to select an intervention component to be delivered via a mobile device. In this work, we present a data analysis method for estimating the average of a long-term positive health outcome that would accrue should a given JITAI be followed. We also provide inferential methods for comparing different JITAIs. The proposed method is illustrated by simulation and a real data application on HeartSteps Study, a mobile intervention study for physical activity.

email: *pengliao@umich.edu*

A BAYESIAN APPROACH FOR EXPLORING HETEROGENEOUS TREATMENT EFFECTS AND INDIVIDUALIZED TREATMENT DECISIONS

Nicholas C. Henderson*, Johns Hopkins University

Thomas A. Louis, Johns Hopkins University

Ravi Varadhan, Johns Hopkins University

Individuals often respond differently to identical treatments, and characterizing such variability in treatment response is an important aim in the practice of personalized medicine. In this talk, we describe a Bayesian approach for utilizing a potentially large number of patient-specific covariates to investigate and quantify the extent of treatment effect differences across study participants. In addition, we outline how our approach can be used both to identify subgroups of patients that may receive substantial treatment benefit and to guide the development of individualized treatment rules. Finally, we illustrate our proposed methodology with data from a clinical trial examining the treatment of chronic heart failure.

email: *nhenders@stat.wisc.edu*

EARLY PHASE DESIGNS FOR TARGETED AND IMMUNOTHERAPEUTIC AGENTS: PREPARING FOR PRECISION MEDICINE

Cody Chiuzan*, Columbia University

The current explosion of targeted and immunotherapeutic agents for cancer treatment has challenged statisticians to reconsider early-phase designs previously developed for cytotoxic agents. The goal of determining the maximum tolerated dose (MTD) is no longer desirable because novel agents are characterized by a reduced toxicity profile, to the point of being essentially safe within the therapeutic dose range. Moreover, for targeted agents, the relationship between target effect and toxicity might not be linear, implying that the most efficacious dose might be below the MTD. Under these circumstances dose selection should not be based solely on toxicity but also examine both toxicity and activity. Recent phase I trials for single-agent or combination therapy have focused on detecting signals of antitumor activity, pharmacokinetic/pharmacodynamics (PK/PD) relationships, or on assessing feasibility and utility of biological correlative assays. Progress has been made, yet there is still much confusion about how and which biological endpoints to incorporate for determining the optimal dose and/or schedule of new drugs for phase II trials. We provide an illustrative review of the most recent early-phase designs proposed for targeted and immunotherapy agents, and discuss their applicability and challenges in the context of the new precision medicine strategy.

email: *cc3780@cumc.columbia.edu*

ESTIMATING OPTIMAL TREATMENT RECOMMENDATION IN OBSERVATION STUDIES

Haoda Fu, Eli Lilly and Company

Nan Jia*, Eli Lilly and Company

With new treatments and novel technology available, personalized medicine has become an important piece in the new era of medical product development. Traditional statistics methods for personalized medicine and subgroup identification primarily focus on single treatment or two arm randomized control trials. Motivated by the recent development of outcome weighted learning framework, we developed algorithms to search treatment assignment rules applying to observational studies. The performance is evaluated by simulations, and we apply our method to a dataset from a diabetes study.

email: *jia_nan2@lilly.com*

COMBINING FUNCTIONAL ADDITIVE MODELS AND ADVANTAGE LEARNING FOR ESTIMATING A TREATMENT DECISION RULE

Adam Ciarleglio*, New York University School of Medicine

Eva Petkova, New York University School of Medicine

R. Todd Ogden, Columbia University

Thaddeus Tarpey, Wright State University

Major depressive disorder (MDD) is a disease characterized by substantial heterogeneity in response to treatment: what works for one patient may be ineffective or harmful for another. This makes treatment selection a difficult task particularly because there are no widely accepted biomarkers for MDD treatment response. Recently, the search for such biomarkers has broadened to include measures derived from neuroimaging modalities (e.g., MRI, fMRI, and EEG) that can be collected before treatment begins. This seems justified since various aspects of brain structure and function have been implicated in depressive symptoms and in response to treatment. We propose an approach for using these functional imaging data to both select the "best" MDD treatment for the individual and provide interpretable measures of the relationship between the imaging data and treatment response. Our approach combines the flexibility of functional additive models with advantage learning in order to obtain treatment decision rules. The approach is evaluated in several realistic settings using synthetic data and is applied to real data arising from a multi-center clinical trial comparing two treatments for MDD in which baseline imaging data are available for subjects who are subsequently treated.

email: *Adam.Ciarleglio@nyumc.org*

COMPANION DIAGNOSTIC DEVICE PARTIAL BRIDGING STUDY IN PRECISION MEDICINE - CHALLENGES AND METHODS

Meijuan Li, U.S. Food and Drug Administration

Yaji Xu*, U.S. Food and Drug Administration

Applications of personalized medicine are becoming increasingly prominent. A well-characterized market-ready companion diagnostic assay (CDx) is often desired for patient enrollment in device-drug pivotal clinical trial(s) so that Food and Drug Administration can ensure that appropriate clinical and analytical validation studies are planned and carried out for CDx. However, such a requirement may be difficult or impractical to accomplish. A clinical trial assay (CTA) instead of CDx may be used for patient enrollment in the clinical trial. However, during the course of clinical trial, CDx may be available as such CTA will be switched to CDx as the clinical trial enrollment assay. A so-called partial bridging study is needed to assess the agreement between CDx and CTA in order to bridge the clinical data from CTA to CDx. In this paper, we will discuss statistical challenges in study design and data analysis for the partial bridging study. Particularly, we aimed to provide statistical methods on how to estimate the drug efficacy in CDx intended use population using results from the partial bridging study and pivotal clinical trial.

email: *yaji.xu@fda.hhs.gov*

IDENTIFYING PREDICTIVE MARKERS FOR PERSONALIZED TREATMENT SELECTION

Yuanyuan Shen*, Harvard School of Public Health

Tianxi Cai, Harvard School of Public Health

It is now well recognized that the effectiveness and potential risk of a treatment often vary by patient subgroups. Although trial-and-error and one-size-fits-all approaches to treatment selection remains a common practice, much recent focus has been placed on individualized treatment selection based on patient information. Genetic and molecular markers are becoming increasingly available to guide treatment selection for various diseases including HIV and breast cancer. In recent years, many statistical procedures for developing individualized treatment rules (ITRs) have been proposed. However, less focus has been given to efficient selection of predictive biomarkers for treatment selection. The standard Wald test for interactions between treatment and the set of markers of interest may not work well when the marker effects are non-linear. Furthermore, interaction based test is scale dependent and may fail to capture markers useful for predicting individualized treatment dierences. In this paper, we propose to overcome these difficulties by developing a kernel machine (KM) score test that can efficiently identify markers predictive of treatment difference. Simulation studies show that our proposed KM based score test is more powerful than the Wald test when there is non-linear effect among the predictors and when the outcome is binary with non-linear link functions. Furthermore, when there is high-correlation among predictors and when the number of predictors is not small, our method also over-performs Wald test. The proposed method is illustrated with two randomized clinical trials.

email: *crystalshenyuany@gmail.com*

132. SURVIVAL ANALYSIS

A UNIFIED SLICE SAMPLER FOR REGRESSION ANALYSIS OF CURRENT STATUS DATA UNDER LINEAR TRANSFORMATION MODELS

Sheng-Yang (Sean) Wang*, University of South Carolina

Lianming Wang, University of South Carolina

Linear transformation models (LTMs) are a broad class of semi-parametric regression models treating the proportional hazards model, proportional odds model and probit model as special cases. Although LTMs are widely used for analyzing right-censored survival data in the literature, their applications for current status data and general interval-censored data are limited. This paper proposes a unified Bayesian estimation approach for regression

analysis of current status data. Our proposed approach adopts monotone splines for modeling the unknown increasing functions in LTMs and uses shrinkage priors for the spline coefficients to allow spline basis selection and to avoid overfitting. A novel slice sampler is proposed to facilitate the posterior computation and to allow one to estimate baseline function and regression parameters simultaneously. The proposed approach is generic for all LTMs and allows model selection. The method is illustrated through application to an epidemiological study of uterine broids.

email: *wang367@email.sc.edu*

AN EXTENDED KAPLAN-MEIER ESTIMATOR FOR TIME TO SUCCESS WITH INFORMATIVE CENSORING

Wei Li*, Astellas Pharma Development

Misun Lee, Astellas Pharma Development

In many clinical trials, the primary endpoint is rate of treatment success which can be analyzed by crude rate as well as time to event approaches. In trials where patients may discontinue treatment or die due to poor treatment response, use of standard survival methods such as Kaplan-Meier and log-rank test yield biased results which may provide a conflicting conclusion from one based on crude rate analysis. When a survival analysis is utilized without careful handling of dropouts, bias can be severe. Methods for time to success must therefore handle such informative censoring mechanisms. In this research, we propose a sophisticated algorithm for handling informative censoring in trials where the primary endpoint is composite (eg, a composite of clinical and mycological responses). We utilize the pool of information observed on components of the composite endpoint and other information such as duration of treatment to assign different missing (or censoring) pattern indexes to patients. Various censoring pattern indexes indicate different levels of likelihood of observing success for the censored patients compared to the patients remaining in the trial. Then we propose an extended Kaplan-Meier estimator that adjusts the probability of event for censored patients by accounting for the censoring patterns.

email: *wei.li3@astellas.com*

SURVIVAL DATA FOR MULTIPLE DISEASES FROM STRATIFIED CASE-COHORT DESIGN

Soyoung Kim*, Medical College of Wisconsin;

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

Donglin Zeng, University of North Carolina at Chapel Hill, Chapel Hill

David J. Couper, University of North Carolina at Chapel Hill, Chapel Hill

When the diseases are not rare or the number of cases is large in biomedical studies, measuring expensive covariates from all cases is not feasible due to finance constraint. Under the situation, the generalized case-cohort design was proposed, which is an efficient tool to study the effect of exposure information. In this design, expensive exposure information is collected from subjects only in a random sample, called the subcohort as well as a portion of cases. When it is of interest to study the effect of one risk factor on multiple diseases, Kim et al. (2013) proposed practically efficient estimators for rare diseases by using exposure information for other diseases. In this paper, we extend the estimators in Kim et al. (2013) to generalized stratified case-cohort design for multivariate failure time and seek to find optimal weight. Under this study design, we propose estimating equations with an optimal weight which is able to use extra information for other diseases outside the subcohort. We develop asymptotic properties of the proposed estimators and show that our proposed methods gain efficiency over existing methods based on simulation results. We apply our proposed methods to data from the Atherosclerosis Risk in Communities study.

email: *skim@mcw.edu*

IMPROVED ESTIMATION OF RELATIVE RISK UNDER SMALL SAMPLES USING A GENERALIZED LOG-RANK STATISTIC

Rengyi Xu*, University of Pennsylvania

Pamela A. Shaw, University of Pennsylvania

Devan V. Mehrotra, Merck

When comparing survival times between groups in the setting of proportional hazards, the Cox model is usually used for estimation, and the log-rank test is used for hypothesis testing. However, traditional methods are large samples methods. Mehrotra and Roth introduced a statistic based on the generalized log-rank (GLR) for better efficiency in estimating relative risk in small samples. In this paper, we propose a refined GLR (RGLR) statistic that further improves efficiency in the estimation. We demonstrate in numerical studies across a variety of scenarios that when the sample size is small (i.e., fewer than 40 subjects per group), our RGLR statistic provides a percentage gain in relative efficiency, ranging from 5% to 90%, than both the parametric and Cox model. Furthermore, the RGLR statistic matches or improves the original GLR in all data settings. With respect to hypothesis testing, the RGLR statistic preserves the type I error, while methods based on parametric and Cox model tend to provide an inflated type I error in small samples. We further show that the performance of the parametric model can be influenced by misspecification of the true underlying distribution, while the RGLR approach provides a consistently high relative efficiency and low bias. Our method is illustrated further with application in a real data example.

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

LIFE EXPECTANCY ESTIMATION BASED ON GOMPERTZ FUNCTION

Zugui Zhang*, Christiana Care Health System

Paul Kolm, Christiana Care Health System

In clinical trials and observational studies, life expectancy cannot be estimated from the trial dataset, since survivals over the period of follow-up within these studies are too high. We propose a method to utilize external databases to estimate life expectancy. From these dataset, the Kaplan-Meier estimates will be used to calculate the overall observed survival curve; and Cox-proportional hazard modeling will be applied to investigate hazard ratios for risk factors, including demographic factors, clinical factors and events which occur during the trial. Then we will be able to create a survival curve by applying the Cox-proportional hazard model to the mean Kaplan-Meier survival for patients in these studies with any given set of covariates. We demonstrated that life expectancy can be estimated from the survival curves by applying a modified Gompertz Function, a specific dynamic population model with exponential growth form, based on the observation that mortality rate tends to increase exponentially with age. Also, we will conduct sensitivity analyses to investigate the robustness of estimated life expectancy with respect to changes in disease epidemiology by considering different scenarios of reduction of mortality rates and incidence rates for cardiovascular events both during and after the trial period.

email: *zzhang@christianacare.org*

DETECTING ASSOCIATIONS BETWEEN MICROBIOME COMPOSITION AND TIME-TO-EVENT OUTCOMES

Anna Plantinga*, University of Washington

Ni Zhao, Fred Hutchinson Cancer Research Center

Michael C. Wu, Fred Hutchinson Cancer Research Center

High-throughput sequencing of 16S rDNA genes allows profiling of the composition of microbial samples, and recent studies have shown associations between microbiome composition and several diseases or conditions. Based on this evidence, microbiome profiling is now being used in laboratory studies and clinical trials, settings that often generate time-to-event data. However, there are no existing methods for testing for associations between microbiome composition and censored outcomes. Therefore, we propose a modified score test for testing these associations. We propose a kernel machine Cox regression framework to model the relationship between a censored time-to-event outcome and microbiome composition, adjusting for other covariates. The kernel matrix is constructed by transforming the standard distance metric, which encodes OTU presence and absence or frequency in a phylogeny-based manner. We construct a modified score test for association testing, which allows the use of microbiome-type kernels while still providing reasonable power and controlling type I error.

email: *aplantin@uw.edu*

SEMIPARAMETRIC STRUCTURAL EQUATION MODELS WITH LATENT VARIABLES FOR RIGHT-CENSORED DATA

Kin Yau Wong*, University of North Carolina, Chapel Hill

Danyu Lin, University of North Carolina, Chapel Hill

Donglin Zeng, University of North Carolina, Chapel Hill

Structural equation modeling is commonly used to capture complex structures of relationships among multiple variables, both latent and observed. In this paper, we propose a general class of structural equation models with a semiparametric component for potentially censored survival times. We consider nonparametric maximum likelihood estimation and devise a combined EM and Newton-Raphson algorithm for its computation. We investigate model identifiability and establish consistency, asymptotic normality, and semiparametric efficiency of the estimators. Finally, we demonstrate the satisfactory performance of the proposed methods through simulation studies and provide application to a motivating cancer study that contains a variety of genomic variables.

email: *alexwky@live.unc.edu*