I. POSTERS: BIOMARKERS

Ia. P-VALUE EVALUATION, VARIABILITY INDEX AND BIOMARKER CATEGORIZATION FOR ADAPTIVELY WEIGHTED FISHER'S META-ANALYSIS METHOD IN OMICS APPLICATIONS

Zhiguang Huo Huo*, *University of Florida*

Shaowu Tang, *Roche Molecular Systems, Inc.*

Yongseok Park, *University of Florida*

George Tseng, *University of Pittsburgh*

K independent studies and to provide better biological interpretation by characterizing which studies contribute to meta-analysis. Currently, AW-Fisher suffers from lack of fast, accurate p-value computation and variability estimate of AW weights. When the number of studies K is large, the 3K - 1 possible differential expression pattern categories can become intractable. In this paper, we apply an importance sampling technique with spline interpolation to increase accuracy and speed of p-value calculation. Using resampling techniques, we propose a variability index for the AW weight estimator and a co-membership matrix to characterize pattern similarities between genes. The co-membership matrix is further used to categorize differentially expressed genes based on their meta-patterns for further biological investigation. The superior performance of the proposed methods is shown in simulations. These methods are also applied to two real applications to demonstrate intriguing biological findings.

✉ zhuo@ufl.edu

Ib. EVALUATING ALTERNATIVE APPROACHES TO BOOTSTRAP ESTIMATION PROCEDURES IN TWO-STAGE GROUP SEQUENTIAL DESIGNS

Sara Biesiadny*, *Rice University*

Nabihah Tayob, *University of Texas MD Anderson Cancer Center*

Two-stage group sequential designs with early termination for futility are important and widely used in EDRN Phase 2 and 3 biomarker studies since they allow us to conserve valuable sample specimens in the case of inadequate biomarker performance. Obtaining unbiased estimates using all the available data in completed studies is not trivial. We have previously developed a nonparametric conditional resampling algorithm to unbiasedly estimate both the combination rule and performance of a biomarker panel in a two-stage group sequential design. Here we explore alternative approaches to the computationally intensive bootstrap procedure to obtain estimates of the distributions of our proposed estimators. This will allow us to utilize simulation studies to explore the operating characteristics and identify optimal designs for evaluating a biomarker panel in a study that includes an interim analysis with termination for futility.

✉ txsarab@gmail.com

Ic. NOVEL QUANTILE APPROACH FOR THE IDENTIFICATION OF BIOMARKERS ASSOCIATED WITH TYPE II DIABETES USING NHANES DATABASE

Hanying Yan*, *Columbia University*

Ying Li, *IBM T. J. Watson Research Center*

Xiaoyu Song, *Icahn School of Medicine at Mount Sanai*

Discovering new uses for approved drugs to provide possible transition from bench to bedside is a crucial strategy to reduce time and cost and improve success rate for drug development. The investigation of hidden connections

between diseases and biomarkers is a critical step to facilitate such drug repurposing research. Existing studies have leveraged the valuable resource of National Health and Nutrition Examination Survey (NHANES) but only considered traditional statistical approaches such as linear models and correlation analyses. In this study, we propose a robust quantile rank-score based approach (QRank) to incorporate the features of complex survey data (multistage stratified, cluster-sampled and unequally weighted sampling scheme). The novel QRank approach is able to identify biomarkers with entire distribution differentially affected by diabetes status. We applied this approach to NHANES 2013-2014 dataset and identified more potential biomarkers which were more likely to be validated in two independent NHANES datasets (2009-2010 and 2011-2012). It suggests that our QRank approach is promising for survey study and the analysis of drug repurposing.

✉ hanying_yan@outlook.com

Id. LEARNING SUBJECT-SPECIFIC DIRECTED ACYCLIC GRAPHS (DAGS) FROM HIGH-DIMENSIONAL BIOMARKER DATA

Shanghong Xie*, *Columbia University*

Xiang Li, *Janssen Research & Development, LLC*

Peter McColgan, *UCL Institute of Neurology, London*

Sarah J. Tabrizi, *UCL Institute of Neurology, London and National Hospital for Neurology and Neurosurgery, London*

Rachael I. Scahill, *UCL Institute of Neurology, London*

Donglin Zeng, *University of North Carolina, Chapel Hill*

Yuanjia Wang, *Columbia University*

The identification of causal relationships between random variables from large-scale observational data using directed acyclic graphs (DAG) is highly challenging. We propose a new mixed-effects structural equation model (mSEM) framework to estimate subject-specific DAGs, where we represent joint distribution of random variables in the DAG as a set of structural causal equations with mixed effects. The directed edges between nodes depend on observed covariates on each of the individual and unobserved latent variables. The strength of the connection is decomposed into a fixed-effect term representing the average causal effect given the covariates and a random effect term representing the latent causal effect due to unobserved pathways. We propose a penalized likelihood-based approach to handle high dimensionality of the DAG model. We theoretically prove the identifiability of mSEM. Extensive simulations and an application to protein signaling data show that the true causal relationships can be recovered from interventional data. We also identify gray matter atrophy networks in regions of brain from patients with Huntington's disease.

✉ sx2168@cumc.columbia.edu

Ie. A REGRESSION-BASED APPROACH INCORPORATING PATIENTS' CHARACTERISTICS TO OPTIMIZE CONTINUOUS DIAGNOSTIC MARKER CUTOFF

Yan Li*, *University of Minnesota*

Chap T. Le, *University of Minnesota*

Many diagnostic markers are either continuous/on an ordinal scale. It is important to dichotomize such markers at an optimum cut-point in order to accurately categorize subjects as diseased or healthy. Le (2006) proposed such a method by maximizing the Youden's Index, which integrates both sensitivity and specificity. However, that approach only consider the relationship between the continuous marker and the disease status, leaving out the subjects' characteristics, which could greatly affect diagnostic-marker cutoff optimization. Here we propose a regression-based model that takes account of patients' characteristics in the dichotomization of the continuous diagnostic biomarker. By analyzing a prostate cancer dataset, we found that the optimal cut-point for acid phosphatase level changes along with subjects' characteristics, including age, cancer stage,

disease grade, and X ray status in the diagnosis of nodal involvement. A comprehensive evaluation using all patients' characteristics in the dichotomizing efforts will be very informative and efficient for clinical research at the era of individualized medicine.

✉ lixxx810@umn.edu

If. gClinBiomarker: AN R PACKAGE FOR CLINICAL BIOMARKER ANALYSES, ALONG WITH "ONE-CLICK" REPORT GENERATING TEMPLATES

Ning Leng*, *Genentech*

Alexey Pronin, *Genentech*

Christina Rabe, *Genentech*

Doug Kelkhoff, *Genentech*

Kwame Okrah, *Genentech*

Jane Fridlyand, *Genentech*

Zhuoye Xu, *Genentech*

Imola Fodor, *Genentech*

Recent development of high throughput technology provides unprecedented power in discovering clinical biomarkers from retrospective analyses. However, these data-driven analyses can also introduce statistical caveats, which may diminish a biomarker's reproducibility in future trials. A number of statistical analyses had been used to sophistically evaluate those potential caveats. We developed an R package gClinBiomarker, which allows users to easily perform such analyses and generate high-quality figures and tables. The gClinBiomarker package contains functions covering essential biomarker analyses conducted in both oncology and non-oncology trials. More importantly, it also provides a series of R markdown templates that allow users to plug in their data and generate a biomarker analysis report by 'one-click'. The report and functions cover selection bias analysis, biomarker property characterization, cutoff exploration (for continuous biomarker), subgroup analysis, ROC analysis, and longitudinal analysis.

✉ leng.ning@gene.com

2. POSTERS: LONGITUDINAL DATA ANALYSIS

2a. PENALIZED SMOOTHING SPLINES IN ADOLESCENT GROWTH STUDIES

Justin M. Leach*, *University of Alabama at Birmingham*

Inmaculada Aban, *University of Alabama at Birmingham*

Adolescent patients with Sickle Cell Anemia (SCA) may demonstrate impaired growth. To better understand the impact of transfusion, hydroxyurea, or no modifying therapy on growth, we retrospectively analyzed the growth patterns over 18 years for a large cohort of patients from one institution with HbSS or SB0 thalassemia. Employing mixed models can account for within subject correlation in longitudinal data analyses, but cases with nonlinear relationships require the assumption of a functional form to have valid inferences. In modeling the relationship between age and growth, a quadratic fit outperforms a linear fit, but is both biologically implausible and a poor fit to the data. Fitting penalized splines to within subject data with the R package fda allows for flexible and more reasonable estimates of growth trajectories. We then chose discrete points from the fitted curves and used the mixed model with age as a categorical variable to avoid further functional assumptions and simplify computation. Alternatively, one may incorporate a functional form into the within-subject portion of the mixed model. We explore the benefits and drawbacks of the two approaches.

✉ jleach@uab.edu

2b. INFERENCE ON MEAN QUALITY ADJUSTED LIFETIME USING JOINT MODELS FOR CONTINUOUS QUALITY OF LIFE PROCESS AND TIME TO EVENT

Xiaotian Gao*, *University of Pittsburgh*

Xinxin Dong, *Takeda Development Center Americas, Inc.*

Chaeryon Kang, *University of Pittsburgh*

Abdus S. Wahed, *University of Pittsburgh*

Quality-adjusted lifetime (QAL) has been considered as an objective measurement that summarizes the quantitative and qualitative health aspects in a unitary and meaningful way. The idea is to account for each patient's health experience adjusting for the overall survival, with death at one extreme and perfect health at the other extreme. In existing literature, the health states are defined to be discrete and the number of states are taken to be finite. Therefore, QAL can be calculated as a sum of time spent at each health state multiplied by the corresponding weight. In this paper we consider how to consistently estimate the mean QAL when the QOL process is assumed to be continuous and observed with error over time at fixed time points via joint modeling. An estimator, together with its asymptotic properties, is presented and investigated through simulation. We then illustrate the method by analyzing a breast cancer clinical trial dataset.

✉ xig31@pitt.edu

2c. A BAYESIAN JOINT FRAILTY-COPULA MODEL FOR RECURRENT EVENTS AND A TERMINAL EVENT

Zheng Li*, *The Pennsylvania State Health Milton S. Hershey Medical Center*

Vernon M. Chinchilli, *The Pennsylvania State Health Milton S. Hershey Medical Center*

Ming Wang, *The Pennsylvania State Health Milton S. Hershey Medical Center*

Recurrent events could be censored by a terminal event, which commonly occurs in biomedical and clinical studies. The non-informative censoring assumption could be violated because of potential dependency between these two event processes. The joint frailty model is one of the widely used approaches to jointly model these two processes; however, several limitations exist: 1) the terminal and recurrent event processes are assumed to be conditionally independent given a subject-level frailty; 2) the association between two processes cannot be directly estimated. In order to fill these gaps, we propose a novel joint frailty-copula approach embedded within a Bayesian framework to model recurrent events and a terminal event; Metropolis-Hastings within the Gibbs Sampler algorithm is used for parameter estimation. We also proposed a method to dynamically predict the time to the terminal event based observed recurrent event. Extensive simulation studies are conducted to evaluate the efficiency and robustness of our proposal. Finally, we apply our method into a real example extracted from the MarketScan database to study the association between recurrent strokes and mortality.

✉ zxl141@psu.edu

2d. SAMPLE SIZE CALCULATION FOR A TWO-GROUP COMPARISON OF REPEATED COUNT OUTCOMES USING GEE WITH NEGATIVE BINOMIAL DISTRIBUTION

Dateng Li*, *Southern Methodist University*

Jing Cao, *Southern Methodist University*

Song Zhang, *University of Texas Southwestern*

Randomized clinical trials with count measurements are common in various medical areas. Controlled clinical trials commonly assign subjects randomly to one of two treatment groups and repeatedly evaluates them at baseline and intervals across a treatment period of a fixed duration. The primary interest is to either compare the rates of change or to compare the time-averaged response (TAD) between treatment groups. Generalized estimating equations (GEEs)

have been widely used because of its robustness to mis-specification of the true correlation structure. Negative binomial regression is commonly used to model count outcome due to its flexibility of taking over-dispersion into consideration. In this paper, we derive sample size formulae for comparing the rates of change or the TAD between two groups in a repeatedly measured count outcome assuming negative binomial distribution and using GEE method. The sample size formula incorporates general missing patterns such as independent missing and monotone missing, and general correlation structures such as AR(1) and compound symmetry. The performance of the sample size formula is evaluated through simulation studies.

✉ datengl@smu.edu

2e. BAYESIAN LONGITUDINAL MULTIPLE OUTCOMES MODELS FOR EXPOSURE EFFECTS

Omar Mbowe*, *University of Rochester*

Edwin van Wijngaarden, *University of Rochester*

Daniel W. Mruzek, *University of Rochester*

Sally W. Thurston, *University of Rochester*

Many birth cohort studies evaluate exposure effects on child development longitudinally, using data from age-specific test batteries administered at multiple ages. Inference from separate models for each outcome cannot estimate overall exposure effects and ignores correlations between outcomes. Longitudinal models are most appropriate when the same outcome is measured over time, but many tests require different instruments for each age, e.g. cognition measures emphasize different skills for preschool versus school-age children. Previous papers examined exposure effects on multiple outcomes at one age in a single model. Extending this, we fit a Bayesian longitudinal multiple outcomes model to examine effects over time on overall cognition (for which each test is assumed to measure a component). We illustrate this with data from the Seychelles Child Development Study, designed to study prenatal methylmercury (MeHg) effects on childhood neurodevelopment. Using data from multiple age-specific outcomes at several ages, we estimate MeHg effects on each outcome and on overall cognition at each age with increased power relative to separate models.

✉ Omar_Mbowe@URMC.Rochester.edu

3. POSTERS: STATISTICAL GENETICS AND GENOMICS

3a. SAME-CLUSTERING: SINGLE-CELL AGGREGATED CLUSTERING VIA MIXTURE MODEL ENSEMBLE

Ruth Huh*, *University of North Carolina, Chapel Hill*

Yuchen Yang, *University of North Carolina, Chapel Hill*

Jin Szatkiewicz, *University of North Carolina, Chapel Hill*

Yun Li, *University of North Carolina, Chapel Hill*

Clustering single-cell RNA-seq (scRNA-seq) data is an important task. Clustering results themselves are of great importance for shedding light on tissue complexity. Inferred cell types from clustering analysis are valuable for differential expression analysis where we need to adjust for them to mitigate deceiving results. Several novel methods have been developed for clustering scRNA-seq data. However, different approaches generate varying cluster assignments and number of clusters. It is usually hard to gauge which method to use because none of the clustering methods always outperform the other methods in all datasets. Our SAME-clustering takes multiple clustering results of a dataset to produce an improved combined clustering. SAME-clustering adopts a probabilistic model to build a consensus by using a finite mixture model of multinomial distributions. Our current implementation takes clustering results from four methods, SC3, CIDR, Seurat, and t-SNE+kmeans, as input and produces ensemble results using the EM algorithm. We have tested SAME-clustering across 14 datasets and results show that our method yields enhanced cluster results.

✉ rhuh@live.unc.edu

3b. INTEGRATING eQTL DATA WITH GWAS SUMMARY STATISTICS IN PATHWAY-BASED ANALYSIS

Chong Wu*, *University of Minnesota*

Wei Pan, *University of Minnesota*

Many genetic variants affect complex traits through gene expression, which can be exploited to boost power and enhance interpretation in GWASs as demonstrated by the transcriptome-wide association study (TWAS) approach. Furth, due to polygenic inheritance, a complex trait is often affected by multiple genes with similar function as annotated in gene pathways. Here we extend TWAS from gene-based analysis to pathway-based analysis. The basic idea is to impute the genetically regulated component of gene expression for each gene in a pathway, then adaptively test for association of the pathway with a trait by effectively aggregating possibly weak association signals across the genes in the pathway. We applied our proposed test with the KEGG and GO pathways to two schizophrenia (SCZ) GWAS summary association datasets, denoted SCZ1 and SCZ2 with about 20,000 and 150,000 subjects respectively. We identified 15 novel pathways associated with SCZ, which could not be uncovered by the single SNP-based analysis or gene-based TWAS. Our results showcase the power of incorporating gene expression information and gene functional annotations into pathway-based association testing for GWAS.

✉ wuxx0845@umn.edu

3c. MethylSeqDesign: A FRAMEWORK FOR METHYL-SEQ GENOME-WIDE POWER CALCULATION AND STUDY DESIGN ISSUES

Peng Liu*, *University of Pittsburgh*

George C. Tseng, *University of Pittsburgh*

In the past decade, bisulfite methylation sequencing (Methyl-Seq) has become the most popular technology to study methylation alterations in a genome-wide scale. To our knowledge, no power calculation and study design method is available for Methyl-Seq. There exists over 28 million possible methylation sites in Methyl-Seq, making full coverage of methylation sites impossible even with very deep sequencing depth. The count data nature in methyl-seq to infer proportion of methylation in a cell population also complicates statistical modeling. Here, we propose a "MethylSeqDesign" framework for power calculation and study design of Methy-Seq by utilizing information from pilot data. Differential methylation analysis is based on beta-binomial model and Wald test statistics transformation. Power calculation is achieved by mixture model fitting of p-values from pilot data and a parametric bootstrap procedure. To circumvent the issue of numerous methylation sites, we focus on inference of pre-specified targeted regions. The method is evaluated by simulation and a chronic lymphocytic leukemia dataset. An R package "MethylSeqDesign" is available on github.

✉ pel67@pitt.edu

3d. SINGLE CELL RNA SEQUENCING COUNT MODELLING AND DIFFERENTIAL EXPRESSION ANALYSIS USING UNIQUE MOLECULAR IDENTIFIER

Wenan Chen*, *St. Jude Children's Research Hospital*

Peer Karmaus, *St. Jude Children's Research Hospital*

Celeste Rosencrance, *St. Jude Children's Research Hospital*

Yan Li, *University of Minnesota*

John Easton, *St. Jude Children's Research Hospital*

Hongbo Chi, *St. Jude Children's Research Hospital*

Gang Wu, *St. Jude Children's Research Hospital*

Xiang Chen, *St. Jude Children's Research Hospital*

Single cell RNA sequencing (scRNA-seq) technology is advancing fast and many protocols and platforms have been developed. So far there are two types of RNA count data provided by current scRNA-seq protocols: read based count and unique molecular identifier (UMI) based count. Based

on analyzing multiple single cell data, we observed huge distribution differences between UMI count and read count. We proposed a model explaining these differences. For UMI count, we concluded that there was no need to model drop-out event with zero inflated models. We also evaluated many differential expression analysis methods in terms of false discovery rate (FDR) and precision-recall curve, including SCDE, MAST. Results show that many packages have inflated FDR when applied to UMI count data and are sub-optimal. In addition, we also analyzed the potential batch effects on differential expression analysis. For real data analysis, we did differential expression analysis between the naïve T cell and memory T cell and between wild type and knockout mouse cells and will discuss our discoveries.

✉ wenan.chen@stjude.org

3e. STATISTICAL METHOD OF GENE SET ANALYSIS FOR SINGLE-CELL RNA-Seq DATA

Di Ran*, *University of Arizona*

Shanshan Zhang, *University of Arizona*

Lingling An, *University of Arizona*

Gene set analysis of single-cell RNA-seq data allows, at a single cell resolution, the quantitative characterization of intratumoral heterogeneity in the transcriptional diversity related to oncogenic signaling or immune response. Current methods typically focus on evaluating the expression of biological pathways or prespecified gene sets between given conditions, e.g., disease versus health. However, due to the unknown (label) status of cell type and the complicated composition of tumor subtypes, the existing methods are either inapplicable to this situation or lack of power in detection. We propose a new statistical method to identify differentially expressed pathways or gene sets based on the variation of diverse expression among cells. There are two major advantages of this method. One is that it can work regardless the label status of cell types is given or not. The other is the capability to detect significant pathways associated with discordant subgroups nested in a given type of cells. Through comprehensive simulation studies and real data analyses, we demonstrate that our approach outperforms existing methods, especially when cell subtypes present.

✉ diran@email.arizona.edu

3f. TWO-SIGMA: A TWO-COMPONENT GENERALIZED LINEAR MIXED MODEL FOR scRNA-Seq ASSOCIATION ANALYSIS

Eric Van Buren*, *University of North Carolina, Chapel Hill*

Yun Li, *University of North Carolina, Chapel Hill*

Ming Hu, *Cleveland Clinic Foundation*

Di Wu, *University of North Carolina, Chapel Hill*

Two key challenges in any analysis of single cell RNA-Seq (scRNA-Seq) data are excess zeros due to "drop-out" events and substantial overdispersion due to stochastic and systematic differences. Association analysis of scRNA-Seq data is further confronted with the possible dependency introduced by measuring multiple single cells from the same sample. Here, we propose TWO-SIGMA, a new TWO-component SInGle cell Model-based Association analysis method. The first component models the drop-out probability with a mixed effects logistic regression model, and the second component models the (conditional) mean read count with a log-linear negative binomial mixed effects regression model. Our approach is novel in that it simultaneously allows for overdispersion, accommodates dependency in both drop-out probability and mean mRNA abundance at the single cell level, leads to improved statistical efficiency, and provides highly interpretable coefficient estimates. Simulation studies and real data analysis show advantages in terms of power gain, type-I error control, and parameter estimation over possible alternative approaches.

✉ edvanburen@gmail.com

3g. DIFFERENTIALLY METHYLATED GENES ASSOCIATED WITH DRUG RESPONSE

Hongyan Xu*, *Augusta University*

Fengjiao Hu, *National Institute of Environmental Health Sciences, National Institutes of Health*

Santu Ghosh, *Augusta University*

Sunil Mathur, *Texas A&M University, Corpus Christi*

Varghese George, *Augusta University*

DNA methylation has long been involved in inter-individual variations in drug response. In this study, we focused on the methylation changes associated with the response in terms of triglyceride changes before and after the treatment with fenofibrate using the real data set. We analyzed genome-wide methylation data from Illumina 500K methylation array. Subjects were categorized into responders and non-responders according to percent changes in triglyceride. We then applied a novel spatial scan statistic to identify genes that are differentially methylated between the responders and non-responders. All the CpG sites within a gene were analyzed together. The spatial scan statistic approach uses a mixed-effects model to incorporate correlations of methylation rates among CpG sites. We analyzed the methylation data at visit 2, accounting for the effects of age, sex, and smoking status as covariates. Methylation levels at 312 genes from 22 autosomes were significantly associated with drug response with $p < 0.01$.

✉ hxu@augusta.edu

3h. INTEGRATED MODELING OF MASSIVE MULTIPLE-DOMAIN DATA

Dongyan Yan*, *University of Missouri*

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

Subharup Guha, *University of Missouri*

Rapid technological advances have allowed for molecular profiling across multiple 'omics domains from a single sample for clinical decision making in many diseases, especially cancer. As tumor development and progression are dynamic biological processes involving composite genomic aberrations, key challenges are to effectively assimilate information across these domains to identify genomic signatures and biological entities that are druggable, develop accurate risk prediction profiles for future patients, and identify novel patient subgroups for tailored therapy and monitoring. We propose an innovative, flexible and scalable Bayesian nonparametric framework for analyzing multi-domain, complexly structured, and high throughput modern array and next generation sequencing-based 'omics datasets. We invent integrative probabilistic frameworks for massive multiple-domain data that coherently incorporate dependence within and between domains to accurately detect tumor subtypes and predict clinical outcomes, thus providing a catalogue of genomic aberrations associated with cancer taxonomy.

✉ dyyr2@mail.missouri.edu

3i. INTEGRATIVE GENE-ENVIRONMENT INTERACTIONS FROM MULTI-DIMENSIONAL OMICS DATA IN CANCER PROGNOSIS

Yinhao Du*, *Kansas State University*

Guotao Chu, *Kansas State University*

Fei Zhou, *Kansas State University*

Cen Wu, *Kansas State University*

Gene-environment interactions have been extensively investigated for their associations with cancer prognosis, where the "gene" refers to a single type of omics features, such as gene expressions or single nucleotide polymorphisms (SNPs). In multi-platform cancer omics studies, multiple types of genomic, genetic and epigenetic changes play important roles in cancer prognosis. In this study, we propose a two-stage penalized variable selection methods for integrative GE interactions from multi-dimensional

omics data. In the first stage, we identify the sparse regulatory effects among multiple types of omics features. In the second stage, we adopt the penalization procedures to identify the GE interactions that are associated with cancer outcomes. The regulated gene expression measurements and their corresponding regulators can be efficiently identified. The advantage of the proposed method has been indicated in extensive simulation studies and a case study on TCGA lung cancer data. The identified interaction effects have important implications in lung cancer prognosis.

✉ ydu0088@gmail.com

3j. eQTL ANALYSIS USING HUMAN RNA-Seq DATA WITH TrecASE AND RASQUAL

Vasyl Zhabotynsky*, *University of North Carolina, Chapel Hill*

Yi-Juan Hu, *Emory University*

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

Wei Sun, *Fred Hutchinson Cancer Research Center*

RNA sequencing at present is a dominant technology to access transcription abundance and gene expression Quantitative Trait Locus (eQTL). Compared with older microarray technology RNA-seq allows one to evaluate allelic imbalance using total expression of the gene as well as allele-specific expression. While allele-specific expression is an invaluable additional source of information and provides more flexibility than older microarrays technology, to fully utilize its benefits it requires careful mapping and allele-specific quantification, which in turn must rely on well imputed data or otherwise account for sequencing bias. We propose a protocol that provide guidelines for data processing. In addition, we compare the underlying assumptions and performance of two top performing eQTL-mapping methods: TrecASE and RASQUAL.

✉ vasyl@unc.edu

4. POSTERS: MISSING DATA AND MEASUREMENT ERROR

4a. AN ESTIMATING EQUATION APPROACH TO ACCOUNTING FOR CASE CONTAMINATION IN EHR-BASED CASE-CONTROL STUDIES

Lu Wang*, *University of Pennsylvania*

Aeron Small, *Yale University*

Rebecca A. Hubbard, *University of Pennsylvania*

Scott M. Damrauer, *University of Pennsylvania*

Jinbo Chen, *University of Pennsylvania*

Clinically relevant information from electronic health records (EHRs) permits derivation of a rich collection of phenotypes. Unfortunately, the true status of any given individual with respect to the trait of interest is not necessarily known. A common study design is to use structured data to identify case and control groups on which subsequent analyses are based. While controls can be identified at high accuracy through rigorous selection criteria, the stringency of rules for identifying cases needs to be balanced against achievable sample size. The inaccurate identification results in a pool of candidate cases consisting of genuine cases and non-case subjects. This case contamination issue represents a unique challenge in EHR-based case-control studies. It is different from the classical case-control misclassification because the non-cases do not satisfy the control definition. We propose a novel estimating equation (EE) approach for estimating odds ratio parameters. Our estimator is consistent and asymptotically normally distributed. We demonstrated our method through extensive simulation studies and an application to a real EHR-based study of aortic stenosis.

✉ luwang6@pennmedicine.upenn.edu

4b. BAYESIAN NONPARAMETRIC ANALYSIS OF LONGITUDINAL DATA WITH NON-IGNORABLE NON-MONOTONE MISSINGNESS

Yu Cao*, *Virginia Commonwealth University*

Nitai Mukhopadhyay, *Virginia Commonwealth University*

Longitudinal data are often infested with missing values. When the missing mechanism is related to the outcome measures, it is non-ignorable and missing not at random (MNAR). When the missing mechanism follows a monotone pattern, the problem of modeling the missing data is simpler and a lot of research has been done on this type of missingness. When the pattern is intermittent, an imputation-based method through an observed-data-missing-value (ODMV) identifying restriction was proposed to estimate the likelihood of the data conditional on missing patterns. However, this model suffers from over-parameterization and is difficult to apply when the sample size is small or the repetitions of outcome measures are large. Therefore, we developed a multi-class missing pattern classification method using neural network with a nonparametric Bayesian framework and a shared-parameter PMM to estimate the conditional likelihood. Simulation studies with varying outcome measures and missing rates were performed showing the advantages of our method with reduced parameters and improved accuracy in dealing with missingness.

✉ caoy4@mymail.vcu.edu

4c. INFORMATIVE DROPOUT WITH NESTED REPEATED MEASURES DATA

Enas Mustfa Ghulam*, *Cincinnati Children's Hospital Medical Center*

Rhonda D. Szczesniak, *Cincinnati Children's Hospital Medical Center*

A frequent problem in medical device studies involve dropout or missing data while collecting multiple long-term follow-up information for each subject. Monitoring measurements at different period of time, such as before and after an intervention, results in nested repeated measures (NRM). Often, it is of interest to examine these measurements at different periods or recording sessions on each individual. Various techniques such as complete case analyses, multiple imputations, and last observation carried forward, were proposed to address dropout in the modeling of longitudinal studies. However, most of these techniques do not take into account the hierarchical structure of the longitudinal data. To overcome this limitation, we propose a joint model to incorporate dropout event and time to dropout in the NRM studies. We examine various imputation techniques through simulation studies. Furthermore, we illustrate these techniques for data collected from a sleep medicine study with parallel groups and NRM; 24-hour ambulatory blood pressure monitoring was taken before and after surgery for each subject.

✉ ghulamem@mail.uc.edu

4d. MULTIPLE IMPUTATION STRATEGIES FOR MISSING CONTINUOUS OUTCOMES IN NON-INFERIORITY RANDOMIZED TRIALS

Lin Taft*, *GlaxoSmithKline*

Douglas Thompson, *GlaxoSmithKline*

Juan Abellan, *GlaxoSmithKline*

Andre Acusta, *GlaxoSmithKline*

Missing data in clinical trials have been widely discussed in the literature, but issues specific to missing data in non-inferiority (NI) trials have rarely been addressed. Very few simulation studies were done to assess different multiple imputation methods to handle missing data specific to NI trials. Here we present the results of a simulation study to examine the properties of various methods of handling missing data to address a hypothetical and a treatment-policy estimand in the context of NI trials where no placebo arm is available. The simulated longitudinal data are generated assuming two active treatment arms with continuous

primary endpoint in 24 scenarios with various treatment discontinuation patterns. For the hypothetical estimand, mixed model for repeated measures is considered with no imputation of missing data. On the other hand, for the treatment-policy estimand, the missing values are imputed with MAR assumption in the control arm, whereas imputed using four MNAR approaches – under the null, Nearest Case, Completer Case, and Average Case in the treatment arm. Operating characteristics are summarized for each method under each scenario.

✉ lin.x.taft@gsk.com

4e. MEASUREMENT ERROR MODELS WITH HIGH-DIMENSIONAL LONGITUDINAL PREDICTORS

Hyung Gyu Park*, *Columbia University*

Seonjoo Lee, *Columbia University*

We propose a new method for estimating multivariate longitudinal measurement error models that use individual intercept and slope from a mixed effect model as predictors. We derive the sufficiency-conditional likelihood of Stefanski & Carroll (1987) that are free of the unobserved "error-prone" subject-specific components, conditioning on their sufficient statistics. The approach was first proposed by Li et al. (2004) for the case of a univariate predictor measured longitudinally, however, we have not seen moderate to high dimensional multivariate longitudinal predictors in measurement error models with appropriate regularizations. The penalized sufficiency-conditional likelihood is optimized via a majorization-minimization coordinate descent algorithm. The method outperforms the naive approach ignoring measurement errors in our simulation examples. This study is motivated by Alzheimer's Disease Neuroimaging Initiative that aims to identify biomarkers that predict dementia transition in elderly participants, in which we utilize longitudinal trajectories of cortical thickness measure as biomarkers.

✉ syhyunpark@gmail.com

4f. COMBINING INVERSE PROBABILITY WEIGHTING AND MULTIPLE IMPUTATION TO ADJUST FOR SELECTION BIAS IN ELECTRONIC HEALTH RECORDS-BASED RESEARCH

Tanayott Thaweethai*, *Harvard University*

David Arterburn, *Kaiser Permanente Washington Health Research Institute*

Sebastien Haneuse, *Harvard University*

While electronic health records (EHR) offer researcher's rich data on large populations over long periods of time, the potential for selection bias is high when analyses are restricted to patients with complete data. Approaching selection bias as a missing data problem, one could apply standard methods. However, these methods generally fail to address the complexity and heterogeneity of EHR data, particularly the interplay of numerous decisions by patients, physicians, and health systems that collectively determine whether complete data is observed. Viewing each such decision as a distinct missingness sub-mechanism, we develop a flexible and scalable framework for estimation and inference in EHR-based research settings that blends inverse probability weighting and multiple imputation. In doing so, one can better align the consideration of missingness assumptions and the analysis to the complexity of the EHR data. The proposed framework is illustrated using data from DURABLE, an EHR-based study of long-term diabetes outcomes following bariatric surgery.

✉ tthaweethai@g.harvard.edu

5. POSTERS: GENOME-WIDE ASSOCIATION STUDIES

5a. APPROXIMATE CONDITIONAL TRAIT ANALYSIS BASED ON MARGINAL GWAS SUMMARY STATISTICS

Peitao Wu*, *Boston University School of Public Health*

Biqi Wang, *Boston University School of Public Health*

James B. Meigs, *Massachusetts General Hospital, Harvard Medical School and Broad Institute*

Josée Dupuis, *Boston University School of Public Health*

Because single genetic variants may have pleiotropic effects, one trait can be a confounder in a genome-wide association study (GWAS) with another trait. The usual way to address this issue is to rerun the analysis adjusting for the confounder. However, it is very time-consuming for a large multi-cohort consortium. We propose an approximate conditional trait analysis based on marginal GWAS summary statistics for two traits, minor allele frequency of the variant, and variances and covariance of two traits. The first three parameters are taken from GWAS meta-analysis while the other parameters are estimated by two strategies: (i) estimates from partial trait data (ii) estimates from published literature. We compare strategies with the estimates based on individual level data. A simulation study for both binary and continuous traits demonstrated good performances of our approach. We also applied our method to the Framingham Heart Study (FHS) GWAS analysis, in which fasting insulin was the trait of interest and body mass index was the confounding trait. A high consistency of genetic effect size was found between our approximate method and individual level data analysis.

✉ peitaowu@bu.edu

5b. A NEW APPROACH FOR EVALUATING THE GLOBAL IMPACT OF MUTATIONS ON GENETIC NETWORKS

Mengqi Zhang*, *Duke University*

Andrew S. Allen, *Duke University*

Complex diseases present challenges to researchers who are trying to decode their genetic mechanisms, which involves a set of disease-related genes but each of them explains a small share of the disease. Recently, network-based analysis integrates signals from individual genes and provides a higher level of network-based view for a better understanding of disease mechanisms. We describe a pathway-based statistical approach to evaluate the global impact of mutations. With prior information extracted from previous studies or knowledge as weights for individual hypotheses, our approach detects signal over the candidate pathway (or other collection of hypotheses) via a weighted Higher Criticism (HC) statistic. It evaluates disease-related networks based on statistical testing results of each gene. This method is specially designed for the situation with large dimensions of data with sparse signals, where a large number of variances with each explains little about the complex disease.

✉ mengqi.zhang@duke.edu

5c. ESTIMATION OF COMPLEX EFFECT-SIZE DISTRIBUTIONS USING SUMMARY-LEVEL STATISTICS FROM GENOME-WIDE ASSOCIATION STUDIES

Yan Zhang*, *Johns Hopkins University*

Guanghao Qi, *Johns Hopkins University*

Juhyun Park, *Dongguk University*

Nilanjan Chatterjee, *Johns Hopkins University*

Summary-level statistics from genome-wide association studies are now widely used to estimate heritability and co-heritability of traits using the popular linkage-disequilibrium (LD) score regression method. We develop a likelihood-based approach for analyzing summary-level

statistics and external LD information to estimate common variants effect-size distributions, characterized by proportion of underlying susceptibility SNPs and a flexible normal-mixture model for their effects. Analysis of summary-level results across 32 GWAS reveals that there is wide diversity in the degrees of polygenicity. The effect-size distributions for susceptibility SNPs could be adequately modeled by a single normal distribution for traits related to mental health and ability and by a mixture of two normal distributions for all other traits. We predict the sample sizes needed to identify SNPs which explain 80% of GWAS heritability to be between 300K-500K for the early growth traits, between 1-2 million for anthropometric and cholesterol traits, between 200K-400K for inflammatory bowel diseases, close to 1 million for chronic diseases and between 1-2 million for psychiatric diseases.

✉ yzhan284@jhu.edu

5d. INTEGRATING GENETIC, TRANSCRIPTIONAL AND BIOLOGICAL INFORMATION PROVIDES INSIGHTS INTO OBESITY: THE FRAMINGHAM HEART STUDY

Jeremiah Perez*, *Boston University*

Lan Wang, *Boston University*

Nancy Heard-Costa, *Boston University*

Audrey Y. Chu, *National Heart, Lung, and Blood Institute, National Institutes of Health*

Roby Joehanes, *Harvard Medical School*

Daniel Levy, *National Heart, Lung, and Blood Institute, National Institutes of Health*

Indices of body fat distribution are heritable, but few genetic signals have been reported from GWAS for CT imaging measurements of body fat distribution. We aimed to identify genes associated with body fat distribution by analyzing gene transcript expression data in blood from participants in the Framingham Heart Study, a large community-based cohort (n up to 4,303), Gene expression signatures at the single gene level were identified for several adiposity related traits, including body mass index, waist-hip ratio and CT-measured indices in men, women, and both sexes combined. Sex-specific effect sizes were also tested. To explore gene regulatory networks, gene expression signatures were identified at the co-expression network module level in the sex-combined sample. By integrating transcriptomic data with results from genomic studies and biological databases genome wide, we identified six gene sets and several key drivers/genes that are central to the adiposity regulatory network. These findings provide a list of compelling candidates for further follow-up studies to uncover biological mechanisms underlying obesity.

✉ jperez993@gmail.com

5e. PCA BASED ADAPTIVE MULTI-TRAIT SNP-SET ASSOCIATION TEST USING THE GWAS SUMMARY STATISTICS

Bin Guo*, *University of Minnesota*

Baolin Wu, *University of Minnesota*

In the past decade, GWAS have identified tens of thousands of genetic variants associated with various diseases and traits. However, there are substantial missing heritability with many genetic variants remaining to be discovered. This is mainly due to the small effect sizes of most genetic variants, and polygenic nature of most traits. In this paper, we develop novel multi-trait multi-variant joint association test methods with improved power by leveraging multiple correlated traits and integrating multiple variants. Our proposed methods have several novel features: (1) just need GWAS summary data; and (2) properly estimate and account for the dependence among traits and variants without raw data to accurately control type I errors; and (3) computationally are scalable to genome-wide association test without the need of resampling or permutation. We illustrate the utility of our proposed methods through rigorous numerical studies and application to analysis of GWAS summary data for multiple lipids traits. Our approach identified many novel loci that have been missed by the current individual trait based tests.

✉ guoxx617@umn.edu

6. POSTERS: SURVIVAL METHODS

6a. SEMI-PARAMETRIC MULTISTATE MARKOV MODEL WITH TRANSITION-SPECIFIC FRAILTY FOR INTERVAL CENSORED LIFE HISTORY DATA

Daewoo Pak*, *Michigan State University*

Chenxi Li, *Michigan State University*

David Todem, *Michigan State University*

We propose a semi-parametric multi-state Markov model having the transition-specific frailty for interval-censored caries life course data. This model is designed for longitudinal investigations on the progression of early childhood caries at the tooth level, and eventually for predicting the probability of caries transition, which can be used for Individualized dental plans for children. We postulate a Cox-type proportional hazards for the transition intensities and use a subject-level bivariate frailty to take into account the intra-oral and inter-transition correlations. The unknown log-baseline intensity is approximated by linear splines. Due to an overfitting problem, the model estimation is conducted using a penalized likelihood through a mixed-model representation. We also develop a Bayesian approach to predict tooth-level caries transition probabilities. The intensive simulation studies for model estimation and prediction indicate that the methods perform well in realistic samples. The practical utility of our model is illustrated using a unique longitudinal study on oral health among children from low-income families in the city of Detroit, Michigan.

✉ pakdaewo@stt.msu.edu

6b. SURVIVAL ANALYSIS UNDER DEPENDENT TRUNCATION

Lior Rennert*, *University of Pennsylvania*

Sharon X. Xie, *University of Pennsylvania*

Traditional methods for truncated survival data rest on the assumption that the survival times are independent of the truncation times. This is not always reasonable in practice. For example, data collected from retirement homes are only obtained from individuals who live long enough to enter the retirement home. Thus, the survival times of these individuals are left truncated by the age at which they enter the home. If individuals who enter the retirement home receive better medical care and live longer, then a dependence structure between the survival and truncation times exists. A violation of the independence assumption can lead to biased estimates. We explore the bias of the hazard ratio estimators from the Cox regression model when the survival and truncation times are dependent. While some methods exist to adjust for dependence, they make stringent parametric assumptions (e.g. copulas). We propose a method to adjust for the dependence with less stringent assumptions, while yielding consistent hazard ratio estimators. We illustrate our approach using an Alzheimer's disease data set, where the truncation and survival times are dependent.

✉ lior.rennert@gmail.com

6c. DIAGNOSTIC ACCURACY ANALYSIS FOR ORDINAL COMPETING RISK OUTCOMES USING ROC SURFACE

Song Zhang*, *University of Pittsburgh*

Many medical conditions are marked by a sequence of events or statuses associated with continuous changes in some biomarkers. Existing methods usually focus on a single cause and compare it with the event-free controls at each time. In our study, we extend the concept of ROC surface and the associated volume under the ROC surface (VUS) from multi-category outcomes to ordinal competing risks outcomes. We propose two methods to estimate the VUS. One views VUS as a numerous metric of correct classification probabilities representing the distributions of the diagnostic marker given the subjects who have experienced different cause-specific events. The other measures the concordance between the marker and the sequential competing outcomes. Since data are often subject to loss of follow up, inverse probability of censoring weight is

introduced to handle the missing disease status due to independent censoring. Asymptotic results are derived using counting process techniques and U-statistics theory. Practical performances of the proposed estimators in finite samples are evaluated through simulation studies.

✉ soz1@pitt.edu

6d. EMPIRICAL LIKELIHOOD INFERENCE FOR SEMIPARAMETRIC TRANSFORMATION MODELS WITH LENGTH-BIASED SAMPLING

Xue Yu*, *Georgia State University*

Yichuan Zhao, *Georgia State University*

The linear transformation models are defined in Chen et al. (2002), which include a general class of semiparametric regression models. Well known proportional hazards model and proportional odds model are special cases. Length-biased data are left-truncated and right censored data under the stationary assumption. In this paper, we propose empirical likelihood and adjusted empirical likelihood inferences for semiparametric transformation models with length-biased sampling, and henceforth showed that under certain regularity conditions, the empirical log-likelihood ratio test statistic converges to a standard chi-squared distribution. Statistical inferences for the regression parameters of interest are made based on the results. Extensive simulation studies are subsequently carried out, and analyzed using real data to illustrate the proposed empirical likelihood and adjusted empirical likelihood methods.

✉ xyu6@gsu.edu

6e. RANDOM FORESTS BASED APPROACH FOR PREDICTION OF COMPETING RISKS UNDER MISSING CAUSE OF FAILURE

Jun Park*, *Indiana University School of Public Health and School of Medicine*

Giorgos Bakoyannis, *Indiana University School of Public Health and School of Medicine*

Ying Zhang, *Indiana University School of Public Health and School of Medicine*

Constantin T. Yiannoutsos, *Indiana University School of Public Health and School of Medicine*

Risk prediction is an important task in modern clinical science and medical decision making. However, the validity of risk prediction based on cohort studies with competing risks and partially observed cause of failure is in general questionable due to missingness. Parametric approaches, such as multiple imputation and augmented inverse probability weighting, under the missing at random assumption, have been proposed for the problem of competing risks with missing cause of failure. However, these approaches have been mainly studied in the context of estimation of regression coefficients and not for risk prediction purposes. They also impose some correct specification assumption regarding the parametric models involved for dealing with missingness. In order to avoid such assumptions, we use a non-parametric imputation approach based on random forests, and rely on a semiparametric competing risks model for risk prediction. We evaluate the performance of this approach via extensive simulations and illustrate the approach using HIV data from a large cohort study in sub-Saharan Africa, where cause of failure is missing for a significant portion of study participants.

✉ jp84@iu.edu

6f. IMPACT OF NON-TERMINAL EVENT STATUS ON HAZARD OF TERMINAL EVENT IN SEMI-COMPETING RISKS DATA

Jing Li*, *Indiana University School of Public Health*

Ying Zhang, *Indiana University School of Public Health*

Giorgos Bakoyannis, *Indiana University School of Public Health*

Semi-competing risks data are a variant of competing risks data. It occurs when a non-terminal event can be censored by a terminal event, but not vice versa. The illness-death model has been widely studied by researchers, which can

be used to characterize the semi-competing risks data. In this project, we choose to incorporate a shared frailty and consider a Markov model when modeling the hazards to study the impact of the illness incidence on mortality in the illness-death model motivated by the Alzheimer's disease data. We are interested in estimating all three hazard functions, healthy status to non-terminal event, and healthy status to the terminal event with and without experiencing the non-terminal event. We propose to estimate the piecewise constant functions for hazards under EM algorithm. We also investigate if the hazard to terminal event changes after the occurrence of non-terminal event with likelihood ratio, score, and Wald tests. Their performance is compared through extensive simulation studies, and the method is applied to a study of Alzheimer disease to illustrate how the incidence of dementia impacts the mortality.

✉ JL204@iu.edu

6g. BAYESIAN ANALYSIS OF SURVIVAL DATA WITH MISSING CENSORING INDICATORS

Veronica J. Bunn*, *Florida State University*

Debajyoti Sinha, *Florida State University*

Naomi Brownstein, *Florida State University*

In some large clinical studies, it may be impractical to give physical examinations to all subjects at every time-point in order to diagnose the occurrence of an event of interest. This challenge creates survival data with missing censoring indicators where the probability of missingness may depend on survival time. We present a fully Bayesian semi-parametric method for such survival data to estimate regression parameters of Cox's proportional hazards model. Simulation studies show that our method performs better than competing methods. We apply the proposed method to data from the Orofacial Pain: Prospective Evaluation and Risk Assessment (OPPERA) study.

✉ v.bunn@stat.fsu.edu

6h. STATISTICAL INFERENCE ON THE CURE TIME VIA CONDITIONAL SURVIVAL

Yueh Wang*, *National Taiwan University*

Hung Hung, *National Taiwan University*

In population-based cancer survival analysis, the net survival is important for government to assess health care programs. Recently, conditional survival is also widely used in clinicians and government to report the survival experience for patients who have already lived certain years. For decades, it is observed that the net survival reaches a plateau after long-term follow-up, this is so called "statistical cure." Several methods were proposed to address the statistical cure. However, those proposed methods assume the cure time to be infinity, thus it is inconvenient to make inference on the cure time. In this study, we define a more general concept of statistical cure via conditional survival. Based on the newly defined statistical cure, the cure time does not necessary to be infinity, We further derive several properties that are helpful to approach the cure rate and cure time.

✉ d02849004@ntu.edu.tw

6i. CHALLENGES AND PITFALLS IN TIME-DEPENDENT SURVIVAL ANALYSIS

Abigail R. Smith*, *Arbor Research Collaborative for Health*

Qian Liu, *Arbor Research Collaborative for Health*

Margaret E. Helmuth, *Arbor Research Collaborative for Health*

Alan B. Leichtman, *Arbor Research Collaborative for Health*

Jarcy Zee, *Arbor Research Collaborative for Health*

Longitudinal covariates are often used to predict survival outcomes; however, it is common for these covariates to be sparsely measured or subject to missingness in retrospective observational studies. Several methods are frequently applied in clinical literature: (1) time-dependent models that remove periods of missing data, (2) time-dependent models that assume the covariate value was unchanged until the next observed value, or (3) time-independent models that average covariate values over the follow-up period and use the average as a baseline covariate. While method (1) can result in extensive inefficiency, methods (2) and (3) are susceptible to bias due to potentially informative missingness and using future information to predict future events, respectively. Balancing the bias and efficiency trade-off in specific applications is challenging, and the magnitude of bias/efficiency loss has not been quantified. Using simulation we compared bias and efficiency of the three methods for estimating the effect of a sparse, binary, time-varying covariate on a time-to-event outcome. We also reviewed results from analyzing prescription drug event data using these various methods.

✉ abby.smith@arborresearch.org

6j. METHOD FOR EVALUATING LONGITUDINAL FOLLOW-UP FREQUENCY: APPLICATION TO DEMENTIA RESEARCH

Leah H. Suttner*, *University of Pennsylvania*

Sharon X. Xie, *University of Pennsylvania*

Current practice of longitudinal follow-up frequency in outpatient clinical research is mainly based on experience, tradition, and availability of resources. Previous methods for designing follow-up times require parametric assumptions about the hazards for an event. There is a need to develop robust, easy to implement, quantitative procedures for justifying the appropriateness of follow-up frequency. We propose a novel method to evaluate follow-up frequency by assessing the impact of ignoring interval-censoring in longitudinal studies. Specifically, we evaluate the bias in estimating hazard ratios using Cox models under various follow-up schedules. Our simulation-based procedure applies the schedules to generated data resembling the survival curve of historical data. Using this method, we evaluate the current follow-up of Parkinson's disease (PD) patients at the University of Pennsylvania Morris K. Udall Parkinson's disease Research Center. However, the method can be applied to any research area with sufficient historical data for appropriate data generation. To allow clinical investigators to implement this method, we provide a Shiny web application.

✉ lsutt@pennmedicine.upenn.edu

6k. OPTIMAL BAYESIAN ESTIMATES OF INVERSE WEIBULL PROGRESSIVE TYPE-II CENSORED DATA

Sarbesh R. Pandeya*, *Georgia Southern University*

Hani Samawi, *Georgia Southern University*

Xinyan Zhang, *Georgia Southern University*

Inverse-Weibull distribution is significantly used to model failure rates in different studies to measure reliability and validity. In this study, we explored the Bayesian inference of inverse Weibull distribution in progressive type-II censored data using both conjugate and Jeffrey's prior distribution. Under this proposed setup, we derived a collection of estimates by using various sets of loss functions in the model. The comparisons for each of the Bayes estimates were made via Monte Carlo simulations. We computed the bias and the mean squared error in order to assess the efficiency of each estimate.

✉ sp03459@georgiasouthern.edu

7. POSTERS: EPIDEMIOLOGICAL METHODS AND CAUSAL INFERENCE

7a. IMPACTS OF GESTATIONAL AGE UNCERTAINTY IN ESTIMATING ASSOCIATIONS BETWEEN PRETERM BIRTH AND AMBIENT AIR POLLUTION

Benjamin E. Nealy*, *Emory University*

Howard H. Chang, *Emory University*

Joshua L. Warren, *Yale University*

Lyndsey A. Darrow, *University of Nevada, Reno*

Matthew J. Strickland, *University of Nevada, Reno*

Previous epidemiologic studies using birth records have shown heterogeneous relationships between air pollution exposure during pregnancy and the risks of preterm birth (PTB, $<$ 37 weeks). Uncertainty in gestational age at birth (GAB) is a previously unquantified variability that may contribute to this heterogeneity. We used logistic regression to identify demographic factors associated with disagreement between clinical and last menstrual period-based (LMP) diagnoses of PTB from individual-level birth certificate data for the Atlanta metropolitan area from 2004 to 2006. We then evaluated the sensitivity in estimated associations between five pollutants during the 1st and 2nd trimesters and PTB when different methods for determining GAB were used. Finally, using a multiple imputation approach, we incorporated uncertainty in the PTB outcome via a binomial distribution with probabilities defined by the proportion of the range between of each birth's clinical and LMP GAB estimates considered preterm. Overall, our analyses demonstrated robust positive associations between PTB and ambient air pollution exposures.

✉ ben.nealy@emory.edu

7b. A SPATIAL FACTOR MODEL APPROACH FOR ASSESSING THE OPIOID EPIDEMIC IN OHIO

David M. Kline*, *The Ohio State University*

Staci A. Hepler, *Wake Forest University*

Andrea E. Bonny, *Nationwide Children's Hospital*

Erin R. McKnight, *Nationwide Children's Hospital*

Opioid misuse is a national epidemic and a significant public health issue due to its high prevalence of associated morbidity and mortality. Ohio has been hit as hard by the opioid epidemic as any state in the country. In Ohio, statewide mental health and addiction services are run through county-level boards overseen by the Department of Mental Health and Addiction Services. Thus, policymakers need estimates of the burden of the opioid problem at the county-level to adequately allocate resources. We propose a joint spatial factor model to assess the burden of opioids as measured by opioid related treatment admission and death counts. We incorporate spatially varying factor loadings to allow for differential contribution to the spatial factor across space. We also estimate effects of county-level social environmental covariates on the spatial factor to better characterize differences in burden between counties. By using more advanced statistical techniques, we can provide policymakers with better information to apply to decision making and resource allocation.

✉ kline.273@osu.edu

7c. A BIOMETRICAL GENETIC MODEL FOR HERITABILITY UNDER ENVIRONMENTAL EXPOSURE OVER MULTIPLE GENERATIONS

Jiali Zhu*, *Kansas State University*

Wei-Wen Hsu, *Kansas State University*

David Todem, *Michigan State University*

Wilfried Karmaus, *University of Memphis*

Polychlorinated biphenyls (PCBs) and dichlorodiphenyldichloroethylene (DDE) are endocrine disrupting chemicals which can imbalance the hormonal system in human body and lead to deleterious diseases such as diabetes, irregular menstrual cycles, endometriosis and breast cancer. These chemicals still exist in the environment and food chains and, once exposed, can be accumulated in human fatty tissues for many years. Moreover, they can be passed from mothers to their offspring through placental transfer or breastfeeding. As a result, the health of the next generation could be potentially affected by the inherited chemicals. In this paper, we investigate how the PCBs and DDE affect the heritability over generations on weight, height and body mass index (BMI) using multigenerational data from the Michigan Fisheater Cohort study. A biometrical genetic model that incorporated multigenerational associations is proposed to investigate the impact of these chemicals on the heritability over generations. Technically, a linear mixed effects model is developed based on the decomposition of phenotypic variance by assuming the environmental effect depends on the maternal exposure levels.

✉ zjl1991@ksu.edu

7d. PROPENSITY SCORE MATCHING WITH MULTILEVEL DATA

Qixing Liang*, *University of Michigan*

Min Zhang, *University of Michigan*

As an alternative to using outcome regression model to adjust for confounding, propensity score (PS) matching method has been a popular method for estimating the causal treatment effect in observational studies. In medical research, often data are of the multilevel structure. Properly adjusting for confounding to estimate the causal treatment effect for multilevel data is more challenging because there may exist both subject-level and cluster-level confounders. Although PS matching has been studied comprehensively for cross-sectional studies, it has not been well studied for multilevel data. We propose a strategy to combine PS matching and outcome regression model for estimating treatment effect while accounting for the hierarchical nature of the data. We show that this method enjoys the double robustness property, i.e. when either PS or outcome model is correctly specified, the bias is negligible. The proposed method has better efficiency and robustness relative to the usual PS method and is more robust than the outcome regression method. Also, it has comparable or better performance than the doubly-robust PS weighted method.

✉ liangqx@umich.edu

7e. ASSESSMENT OF RESIDENTIAL HISTORY AS A SURROGATE FOR ENVIRONMENTAL EXPOSURE

Anny-Claude Joseph*, *Virginia Commonwealth University*

David C. Wheeler, *Virginia Commonwealth University*

In many spatial studies of disease risk, researchers use location at the time of diagnosis as a surrogate for unknown environmental exposures. The implicit assumption is that individuals reside where causal exposures occurred for as long or longer than the latency period of the disease. In the U.S. where residential mobility levels are high, this assumption is questionable. Ignoring mobility may bias estimates of the relationship between the exposure and outcome of interest. To investigate this hypothesis, we conducted a simulation study based on the residential histories of a random sample from the National Institutes of Health American Association of Retired Persons (NIH-AARP) Diet and Health Study. We generated exposure surfaces and assigned exposure and disease status to subjects based on residential histories. We compared estimates from models where exposure was fixed to estimates where exposure varied temporally.

✉ josephac@vcu.edu

7f. EVALUATING SAMPLE SIZES IN COMPARING OVER-DISPERSED COUNT DATA UNDER INCORRECT VARIANCE STRUCTURE

Masataka Igeta*, *Hyogo College of Medicine*

Kunihiko Takahashi, *Nagoya University*

Shigeyuki Matsui, *Nagoya University*

Over-dispersed count data are frequently observed in clinical trials where the primary endpoint is occurrence of clinical events. Sample sizes of comparative clinical trials with these data are typically calculated under negative binomial models or quasi-Poisson models with specified variance functions, or under correct specification of the "working" variance functions. In this presentation, we propose a sample size formula anticipating misspecifications of the working variance function. We derived the formula based on the Wald test statistic with a sandwich-type robust variance estimator under null hypothesis using quasi-Poisson models. The asymptotic variance of the treatment effect estimator involves both true and working variance functions. Our formula includes several existing formulas as special cases when the working variance function is correct. We also consider a sensitivity analysis for possible misspecifications of the "true" variance function. A simulation study demonstrated the adequacy of our formulas. An application to a clinical trial to evaluate the treatment effect on prevention of COPD exacerbation is provided.

✉ masataka.igeta@gmail.com

8. POSTERS: CANCER APPLICATIONS

8a. ESTIMATING LEAD-TIME BIAS IN LUNG CANCER DIAGNOSIS OF CANCER SURVIVORS

Zhiyun Ge*, *University of Texas Southwestern Medical Center and Southern Methodist University*

Daniel Heitjan, *University of Texas Southwestern Medical Center and Southern Methodist University*

David Gerber, *University of Texas Southwestern Medical Center*

Lei Xuan, *University of Texas Southwestern Medical Center*

Sandi Pruitt, *University of Texas Southwestern Medical Center*

Surprisingly, survival from lung cancer has been found to be longer for cancer survivors than for those with no previous cancer. A possible explanation is lead-time bias, which extends survival by advancing the time of diagnosis. We propose a discrete parametric model to jointly describe survival in no-previous-cancer group (no lead-time bias exists) and previous-cancer group (lead-time bias is possible). We model the lead time with a negative binomial distribution and the post-lead-time survival with a linear spline on the logit hazard scale, allowing survival to differ between groups even in the absence of bias. We fit our LS/NB (Logit-Spline/Negative Binomial) model to a propensity-score matched subset of the 2014 Surveillance Epidemiology and End Results (SEER)-Medicare linked data set. For stage I&II lung cancer patients, the estimated mean lead time is roughly 11 months for lung-cancer-specific survival and 3.4 months for overall survival. For higher-stage patients, the lead-time bias is 1 month or less for both outcomes. Accounting for lead-time bias generally reduces the survival advantage of the previous-cancer arm, but it does not nullify it in all cases.

✉ gezhiyun.sunny@gmail.com

8b. CANCER MORTALITY IN USA 1999-2014: A REVIEW AND INTER-STATE COMPARISONS

Desale Habtezgi*, *DePaul University*

Dimitre Stefanov, *University of Akron*

Midha Chand, *University of Akron*

Ashish Das, *Indian Institute of Technology, Bombay*

The United States Center for Disease Control and Prevention (CDC), under the National Program of Cancer Registries, regularly conducts surveys on cancer mortality. The United States Cancer Statistics: 1999–2014 Incidence and Mortality Web-based Report has compiled state-wide cancer incidence and mortality data. In this talk, we use one of the multiple criteria decision making (MCDM) techniques to rank the states based on their 1999–2014 Mortality Data. The cancer cases used for this talk include the five most common cancers in the US, based on the 2013 cancer facts from the American Cancer Society: Female Breast, Colon and Rectum, Pancreatic cancer, Lung and Bronchus, and Prostate.

✉ dhabtzgh@depaul.edu

8c. PROPENSITY SCORE ESTIMATION WITH MISSING COVARIATE DATA

Joanna G. Harton*, *University of Pennsylvania*

Weiwei Feng, *University of Pennsylvania*

Nandita Mitra, *University of Pennsylvania*

Propensity scores are often used in observational studies to estimate treatment effects; however, this approach can be complicated by missing covariate data. We impute the derived propensity score using a passive approach by multiply imputing the missing covariates and then estimating the propensity score. Logistic regression has long been the preferred method for estimating propensity scores, but a misspecified model can result in meaningless estimates. Superlearner, an ensemble machine-learning algorithm that uses cross-validation to evaluate different models, returns a propensity score based on a weighted average but can have inflated variance. The treatment effect estimate can then be obtained by combining the m imputed datasets by either averaging the propensity score for each individual across the datasets (across) or by estimating the treatment effect m times and then averaging those m treatment effect estimates (within). We explore combinations of these approaches using both propensity score matching and inverse-probability treatment weights. We compare bias and efficiency across these various approaches using a National Cancer Database study on head and neck cancer.

✉ jograce@pennmedicine.upenn.edu

8d. STATISTICAL CONSIDERATIONS IN TUMOR-ONLY VARIANT CALLING

Paul L. Little*, *University of North Carolina Lineberger Comprehensive Cancer Center*

David N. Hayes, *University of North Carolina Lineberger Comprehensive Cancer Center, West Cancer Center and University of Tennessee Health Science Center*

Joel Parker, *University of North Carolina Lineberger Comprehensive Cancer Center*

Alan Hoyle, *University of North Carolina Lineberger Comprehensive Cancer Center*

Jose Zevallos, *Washington University School of Medicine, St. Louis*

Heejoon Jo, *University of North Carolina Lineberger Comprehensive Cancer Center*

Angela Mazul, *University of North Carolina Lineberger Comprehensive Cancer Center*

Targeted sequencing and filtering variant calls (VCs) involve sequencing a tumor and matched normal (MN) sample to identify somatic VCs within clinically actionable genes. However, MN sequencing increases marginal costs relative to the amount of genomic data obtained and MNs may not be gathered hindering interpretation of genomic results. With an unmatched normal (UMN) sample as a control, germline VCs and sequencing artifacts contaminate final

VCs. More UMNs lead to increased computational costs and risk of selecting low quality UMNs. We aim to return several dozen confident VCs using a practical number of UMNs. Among our cohort of 1500+ patients from clinical trial LCCC1108:NCT01457196, we selected 100 tumor/MN pairs. Samples were sequenced using Illumina PE sequencing on HiSeq2000/2500 and NextSeq500. Treating these as TO and obtaining a disjoint set of high quality UMNs, variants were called and combined using Strelka, UNCeqR, and Cadabra, annotated with Oncotator, and filtered with our pipeline. At ten UMNs, our VCs achieved an average 95% SENS and 99% SPEC with an average 70 filtered VCs per sample. Our method achieves high SENS/SPEC with a reasonable number of UMNs.

✉ pllittle@email.unc.edu

8e. INTENSITY NORMALIZATION OF MRI IMAGES SPECIFIC TO PATIENTS WITH GLIOBLASTOMAS TO IMPROVE PREDICTION OF TREATMENT OUTCOMES USING RADIOMICS

Abdhi Sarkar*, *University of Pennsylvania*

Russell T. Shinohara, *University of Pennsylvania*

Magnetic Resonance Images (MRIs) of patients with Glioblastomas can be obtained across several visits and sites for different patients on different scanners. For studies investigating patient outcomes for a particular treatment protocol, Radiomics can be effectively used to predict outcomes of a patient. Intensity normalization is a crucial pre-processing step that may significantly influence the accuracy of this prediction. Commonly used techniques such histogram matching and z scoring in the context of tumors may potentially disrupt the pathology of the original image in order to conform to a common atlas or template diluting the signal within the data. We extend a technique called RAVEL that accounts for subject-wise random variation through control voxels to produce biologically interpretable and normalized intensities to facilitate between subject comparisons.

✉ abdhi@pennmedicine.upenn.edu

8f. ASSESSING TIME-DEPENDENT TREATMENT EFFECT WITH JOINT MODEL OF INTERMEDIATE AND TERMINAL EVENTS

Wenjia Wang*, *University of Michigan*

Alexander Tsodikov, *University of Michigan*

In biomedical studies, it is common that patients may experience sequential intermediate and terminal events. Since treatment is assigned based on disease progression, it is informative of prognosis. We consider the situation where treatment is applied at the time of intermediate event, and would like to test its causal effect on the terminal event. A traditional approach regressing time to terminal event on treatment as a time-dependent covariate is misleading when treatment is confounded by indication. To address this problem, we formulate a semiparametric mechanistic joint model and use it to develop a test that is unbiased under the null hypothesis. Profile likelihood and the EM algorithm are used for statistical inference. Large-sample properties of proposed estimators are established. The methodology is illustrated by simulation studies and analysis of real data.

✉ icywang@umich.edu

8g. A MATRIX-BASED APPROACH FOR TWO-GROUP COMPARISON WITH CORRELATED OBSERVATIONS AND HETEROGENEOUS VARIANCE STRUCTURE

Yun Zhang*, *University of Rochester*

Xing Qiu, *University of Rochester*

Two-group comparison is one of the most widely used statistical tools in real life. Student's t-test and Wilcoxon rank-sum test are the standard parametric and nonparametric tests for i.i.d. observations. However, in practice, i.i.d.-ness is hardly met. In this paper, we introduce a rotated test for correlated observations with heterogeneous variance. Based on matrix algebra, we propose to decompose the variance-covariance matrix with general structure (i.e. not equal to the identity matrix) and inversely transform

the observations back to the i.i.d. case. An orthogonal transformation is derived, which uniquely maps the transformed hypothesis testing problem to some standard setting where existing tests can be applied. In simulations, our proposed parametric and nonparametric tests outperform the standard tests in both normal distribution and double exponential distribution. A real data application is demonstrated with RNA-seq data for breast cancer. Our proposed test detects more differentially expressed genes (DEG) than the standard test, and reveals more molecular functions related to the breast cancer.

✉ yun_zhang@urmc.rochester.edu

8h. PREDICTING SURVIVAL TIME IN CANCER PATIENTS USING GENETIC DATA AND IMMUNE CELL COMPOSITION

Licai Huang*, *Fred Hutchinson Cancer Research Center*

Paul Little*, University of North Carolina, Chapel Hill*

Wei Sun, *Fred Hutchinson Cancer Research Center*

Qian Shi, *Mayo Clinic*

Tabitha Harrison, *Fred Hutchinson Cancer Research Center*

Riki Peters*, Fred Hutchinson Cancer Research Center*

Andy Chan, *Massachusetts General Hospital and Harvard Medical School*

Polly Newcomb, *Fred Hutchinson Cancer Research Center*

Previous studies have shown that the types and proportions of tumor infiltrating immune cells have prognostic values in different types of cancer (e.g., Angelova et al. Genome biology 16.1 (2015): 64), and germline genetic variants affect human immune system (Roederer et al. Cell 161.2 (2015): 387-403). Therefore, it is possible that genetic factors affect survival time by modifying the composition of tumor infiltrating immune cells, which can be inferred using gene expression data. However, many large-scale population studies have both genetic and survival time data, but not gene expression data. We propose a two-step approach to identify genetic variants that are associated with survival outcome. We first identify the genetic variants associated with immune cell composition using datasets with both genetic and gene expression data (e.g., The Cancer Genome Atlas, TCGA), and then use such genetic signatures to predict survival time by an elastic net model.

✉ lhuang2@fredhutch.org

9. POSTERS: BAYESIAN METHODS

9a. BAYESIAN PARAMETRIC COVARIANCE REGRESSION ANALYSIS

Guanyu Hu*, *Florida State University*

Fred Huffer, *Florida State University*

Jonathan Bradley*, Florida State University*

This paper introduces Bayesian parametric covariance regression analysis for a response vector. The proposed method defines a regression model between the covariance matrix of a p-dimensional response vector and the auxiliary variables. We proposed constrained Metropolis-Hastings algorithm to get the estimates. Then we demonstrate that the coefficients have posterior consistency when the number of sample size goes infinity or when the dimension of the response goes to infinity. The simulation results are presented to show performance of both regression and covariance matrix estimates. Furthermore, we have more reality simulation experiment to show our Bayesian approach has better performance than MLE in this case. Finally, we have an example of Google Flu data to illustrate the usefulness of our model.

✉ guanyu.hu@stat.fsu.edu

9b. A SEQUENTIAL APPROACH TO BAYESIAN JOINT MODELING OF LONGITUDINAL AND SURVIVAL DATA

Danilo Alvares*, *Harvard School of Public Health*

Carmen Armero, *University of Valencia*

Anabel Forte, *University of Valencia*

Nicolas Chopin, *CREST-ENSAE and HEC Paris*

The statistical analysis of the information generated by medical follow-up is a very important challenge in the field of personalised medicine. As the evolutionary course of a patient's disease progresses, its medical follow-up generates more and more information that should be processed immediately in order to review and update its prognosis and treatment. Hence, we focus on this update process through sequential inference methods for joint models of longitudinal and survival data from a Bayesian perspective. More specifically, we propose the use of sequential Monte Carlo methods for static parameter joint models with the intention of reducing computational time in each update of the inferential process. Our proposal is very general and can be easily applied to most popular joint models approaches. We illustrate the use of the presented sequential methodology in a joint model with competing risk events for a real scenario involving patients on mechanical ventilation in intensive care units.

✉ dalvares@hsph.harvard.edu

9c. COMPARISON OF CONFIDENCE INTERVAL ESTIMATION IN LINEAR EXCESS RELATIVE RISK MODELS BETWEEN BAYESIAN MODEL AVERAGING AND EXPANDED INTERVAL ESTIMATION WHEN EXPOSURE UNCERTAINTY IS COMPLEX

Deukwoo Kwon*, *University of Miami*

Steven L. Simon, *National Cancer Institute, National Institutes of Health*

F. Owen Hoffman, *Oak Ridge Center for Risk Analysis*

In some epidemiologic studies, risk estimation from dose-response analysis can be difficult task since exposure uncertainty is complex. In this situation, conventional regression method shows unsatisfactory performance in interval estimation for risk since it cannot reflect exposure uncertainty adequately. Simon et al. (2015) provided advanced exposure assessment technique to consider complex uncertainty using the two-dimensional Monte Carlo (2DMC) method. From the 2DMC method, multiple exposure realizations can be generated. To date, two approaches to dose response analysis have been published that address situations with multiple sets of exposure realizations: Kwon et al. (2016) and Zhang et al. (2017). The former one is based on Bayesian model averaging (BMA). The latter one is based on conventional regression, increasing the upper portion of the confidence interval using the approximate asymptotic distribution of parameter estimation. In this study, we investigate whether two approaches work well in various scenarios in terms of complexity in exposure uncertainty using simulation study and real data example.

✉ DKwon@med.miami.edu

9d. APPLICATION OF BAYESIAN HIERARCHICAL MODELS FOR ESTIMATING ANNUAL GLOBAL BURDEN OF INFLUENZA-ASSOCIATED HOSPITALIZATION

Guandong Yang*, *Emory University*

Howard H. Chang, *Emory University*

Mingrui Liang, *Emory University*

Katherine M. Roguski, *Centers for Disease Control and Prevention*

Jeremy Reich, *Centers for Disease Control and Prevention*

Neha Patel, *Emory University*

Vanessa Cozza, *World Health Organization, Switzerland*

Julia Fitzner, *World Health Organization, Switzerland*

Danielle A. Iuliano, *Centers for Disease Control and Prevention*

Estimates of influenza-associated hospitalization are important for guiding decisions on prevention measures. Countries without influenza surveillance data may not be able to calculate estimates. We describe an application of a Bayesian hierarchical model to estimate country-specific influenza-associated hospitalizations. We first model age-specific influenza-associated hospitalization rates for 2008-2016 using a meta-analytic framework. To better predict influenza-associated hospitalization for countries without data, we fit a model using available age-specific influenza-associated hospitalization rates as a function of health, economic, and environmental covariates. Example covariates may include influenza transmission zones, gross domestic product, and respiratory infection mortality estimates, which may account for between-country and -year variability in the risk of being hospitalized for an influenza illness episode. Bayesian variable selection addresses potential collinearity among covariates and better accounts for model uncertainty as extrapolations are made by averaging predictions from all possible models with weights determined by the observed data.

✉ yangguandong924@gmail.com

9e. A SPATIALLY VARYING CHANGE POINT MODEL FOR DETERMINING GLAUCOMA PROGRESSION USING VISUAL FIELD DATA

Samuel I. Berchuck*, *University of North Carolina, Chapel Hill*

Joshua L. Warren, *Yale University*

Jean-Claude Mwanza, *University of North Carolina, Chapel Hill*

Early diagnosis of glaucoma progression is critical for limiting irreversible vision loss. A common method for assessing progression uses a longitudinal series of visual fields (VF) acquired at regular intervals. VF progression is defined by slow (or stable) deterioration, followed by a rapid decrease in visual ability. Determining the transition point of the disease trajectory to a more severe state is important clinically for disease management. We introduce a model with spatially varying intercepts, slopes, variances, and change points (CP) to model and predict the trend at each VF location. For each VF location, the change point is a mixture distribution, allowing for both stable and progressing trajectories. The flexible design of the model allows for accurate prediction of the VF data and analyzing the CPs offers a method for determining if each VF location is deteriorating. We show that our new method improves the diagnostic rate of glaucoma progression using data from the Vein Pulsation Study Trial in Glaucoma and the Lions Eye Institute trial registry. Simulations are presented, showing the proposed methodology is preferred over existing models for VF data.

✉ berchuck@live.unc.edu

9f. BAYESIAN BICLUSTERING ANALYSIS VIA ADAPTIVE STRUCTURED SHRINKAGE

Ziyi Li*, *Emory University*

Changgee Chang, *University of Pennsylvania*

Suprateek Kundu, *Emory University*

Qi Long, *University of Pennsylvania*

Biclustering can identify local patterns of a data matrix by clustering rows and columns at the same time. Various biclustering methods have been proposed and successfully applied to analysis of gene expression data. While existing biclustering methods have many desirable features, most of them are developed for continuous data and none of them can handle multiple genomic data types, for example, negative binomial data as in RNA-seq data. In addition, all the existing methods cannot utilize biological information such as those from functional genomics. Recent work has shown that incorporating biological information can improve variable selection and prediction performance in analyses

such as linear regression; In this work, we propose a novel Bayesian biclustering method that can handle multiple data types including Gaussian, binomial, negative binomial, and Poisson data. In addition, our method uses a Bayesian adaptive shrinkage prior that enables feature selection guided by biological information. Simulation studies and application to a cancer genomic dataset demonstrate robust, superior performance of the proposed method, compared to existing biclustering methods.

✉ Ziyi.li@emory.edu

9g. BAYESIAN NONPARAMETRIC FUNCTIONAL MODELS IN MATCHED CASE-CROSSOVER STUDIES

Wenyu Gao*, *Virginia Tech*

Inyoung Kim, *Virginia Tech*

In public health epidemiology, 1-M matched case-crossover studies have become popular. The design compares one case period and M controls on the same patient. The matching covariates from the same patient can be eliminated by the retrospective model when studying the unknown relationship between exposure and the covariates. However, some covariates, such as time and location, play an important role in determining this unknown relationship. Not accounting for these factors can lead to incorrect inferences for all covariates in the model. Hence, in this talk, we propose two functional varying coefficient models to evaluate and assess this unknown relationship. Our approach provides not only the flexibility and interpretability, but also avoids the "curse of dimensionality". We functionally model time-varying coefficients using a regression spline model and an integration model, and we build them under the hierarchical Bayesian framework. We also demonstrate the advantages of our approach using simulation studies together with a 1-4 bidirectional matched case-crossover study using empirical data within public health epidemiology.

✉ wenyu6@vt.edu

9h. BAYESIAN SINGLE INDEX MODEL WITH COVARIATES MISSING AT RANDOM

Kumaresh Dhara*, *Florida State University*

Debdeep Pati, *Texas A&M University*

Debajyoyti Sinha, *Florida State University*

Stuart Lipsitz, *Harvard Medical School*

For many biomedical and environmental studies with unknown non-linear relationship between the response and its multiple predictors, single index model provides practical dimension reduction and good physical interpretation. However widespread uses of existing Bayesian analysis for such models are lacking in biostatistics due to some major impediments including slow mixing of the Markov Chain Monte Carlo (MCMC), inability to deal with missing covariates and a lack of theoretical justification of the rate of convergence. We present a new Bayesian single index model with associated MCMC algorithm that incorporates an efficient Metropolis Hastings step for the conditional distribution of the index vector. Our method leads to a model with good biological interpretation and prediction, implementable Bayesian inference, fast convergence of the MCMC, and a first-time extension to accommodate missing covariates. We also obtain for the first time, the set of sufficient conditions for obtaining optimal rate of convergence of the overall regression function. We illustrate the practical advantages of our method and computational tool via re-analysis of an environmental study.

✉ k.dhara@stat.fsu.edu

9i. A LATENT BAYESIAN CLASSIFICATION MODEL TO PREDICT KIDNEY OBSTRUCTION BASED ON RENOGRAPHY AND EXPERT RATINGS

Changgee Chang*, *University of Pennsylvania*

Jeong Hoon Jang, *Emory University*

Amita Manatunga, *Emory University*

Qi Long, *University of Pennsylvania*

Kidney obstruction is a serious disease which can lead to loss of renal function when not treated in a timely manner. Diuresis renography is widely used to detect obstruction in kidney. However, the diagnosis largely relies on experts' experiences, and there is no gold standard statistical approach designed to analyze renogram curves and clinical variables associated with patients. In this work, we propose an integrative Bayesian approach that models the triplet jointly: renogram curves, clinical variables of patients, and experts' ratings, conditional on the latent kidney obstruction status. In particular, we adopt a nonparametric approach for modeling regnogram curves in which the coefficients of the basis functions are parameterized using latent factors that are dependent on the latent disease status. We develop an MCMC training algorithm and an associated prediction algorithm for kidney obstruction that are computationally efficient. We demonstrate the superior performance of our proposed method in comparison with several naïve approaches via extensive simulations as well as analysis of real data collected from a kidney obstruction study.

✉ changgee@pennmedicine.upenn.edu

9j. BAYESIAN NETWORK META-REGRESSION MODELS USING HEAVY-TAILED MULTIVARIATE RANDOM EFFECTS WITH COVARIATES-DEPENDENT VARIANCES

Hao Li*, *University of Connecticut*

Ming-Hui Chen, *University of Connecticut*

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

Sungduk Kim, *National Cancer Institute, National Institutes of Health*

Arvind K. Shah, *Merck & Co., Inc.*

Jianxin Lin, *Merck & Co., Inc.*

Many clinical trials have been carried out on safety and efficacy evaluation of cholesterol-lowering drugs. To synthesize the results from different clinical trials, we examine treatment level (aggregate) network meta-data from 29 double-blind, randomized, active or placebo-controlled statins +/- Ezetimibe clinical trials on adult treatment-naive patients with primary hypercholesterolemia. In this paper, we assume a multivariate t distribution for the random effects in the proposed network meta-regression model. We further propose a log-linear model for the variances of the random effects so that the variances depend on the aggregate covariates. A variation of deviance information criterion and the logarithm of the pseudo-marginal likelihood based on conditional CPO's are developed for model comparisons. An efficient Metropolis-within-Gibbs sampling algorithm is developed to carry out the posterior computations. We apply the proposed methodology to conduct an in-depth analysis of the network meta-data from 29 trials with 11 treatment arms.

✉ hao.2.li@uconn.edu

9k. BAYESIAN DESIGN OPTIMIZATION OF AN OLFACTORY SENSOR SYSTEM AND A NEW MEASURE OF ANALYTICAL SELECTIVITY

David Han*, *University of Texas, San Antonio*

Using an array of chemical sensors with well calibrated tuning curves, it is possible to appreciate a wide range of stimuli. Using the decision-theoretic approach, inference for the analytes is performed by minimizing the posterior expected loss of choice. Under the same framework, the design optimization of a sensory system is explored via maximization of the expected utility of choice, which suits a particular design aim. Here we formulate the utility functions based on the Shannon information content and the generalized posterior mean squared error, which can be also used as a metric to determine the merit of a sensory system. Furthermore, to characterize the fundamental analytical capability of a measurement system, a general-purpose selectivity is defined in an information-theoretic manner. The proposed metric is not only applicable to systems with linear response functions but also with nonlinear ones. This figure of merit can be also used as an optimality criterion for

the decision-theoretic Bayesian design. Asymptotically, it can be expressed in terms of the Fisher information, and its relation to other notions of selectivity such as the net analyte signal is demonstrated.

✉ david.han@utsa.edu

10. POSTERS: FUNCTIONAL DATA ANALYSIS

10a. SHAPE CONSTRAINED UNIVARIATE DENSITY ESTIMATION

Sutanoy Dasgupta*, *Florida State University*

Debdeep Pati, *Texas A&M University*

Ian Jermyn, *Durham University*

Anuj Srivastava, *Florida State University*

The problem of estimating a pdf with certain shape constraints is both important and challenging. We introduce a geometric approach for estimating pdfs under the constraint that the number of modes of the pdf is fixed. This approach relies on exploring the desired pdf space using an action of the diffeomorphism group that preserves shapes of pdfs. It involves choosing an initial template with desired number of modes and arbitrary heights at the critical points. This template is transformed via composition by a diffeomorphism and subsequent normalization to obtain the final estimate, under the maximum-likelihood criterion. The search for optimal diffeomorphism is accomplished by mapping diffeomorphisms to the tangent space of a Hilbert sphere, a vector space whose elements can be expressed using an orthogonal basis. This framework is applied to shape-constrained univariate, pdf estimation and then extended to conditional pdf estimation. We derive asymptotic convergence rates of the estimator and demonstrate its application on a synthetic dataset.

✉ s.dasgupta@stat.fsu.edu

10b. HYBRID PRINCIPAL COMPONENTS ANALYSIS FOR REGION-REFERENCED LONGITUDINAL FUNCTIONAL EEG DATA

Aaron W. Scheffler*, *University of California, Los Angeles*

Donatello Telesca, *University of California, Los Angeles*

Qian Li, *University of California, Los Angeles*

Catherine Sugar, *University of California, Los Angeles*

Charlotte DiStefano, *University of California, Los Angeles*

Shafali Jeste, *University of California, Los Angeles*

Damla Senturk, *University of California, Los Angeles*

Electroencephalography (EEG) data possess a complex structure that includes regional, functional, and longitudinal dimensions. Our motivating example is a word segmentation paradigm in which typically developing (TD) children and children with Autism Spectrum Disorder (ASD) were exposed to a speech stream. For each subject, continuous EEG signals recorded at each electrode were divided into one-second segments and the power spectral density is estimated. Standard EEG power analyses often collapse information by averaging power across segments and concentrating on frequency bands. We propose a hybrid principal components analysis (HPCA) for region-referenced longitudinal functional EEG data which utilizes vector and functional principal components analyses and does not collapse information along any of the dimensions of the data. The decomposition only assumes weak separability of the higher-dimensional covariance process and utilizes a product of one dimensional eigenvectors and eigenfunctions obtained from the regional, functional, and longitudinal marginal covariances. A mixed effects framework geared towards sparse data structures is proposed to estimate the model components.

✉ ascheffler@ucla.edu

I0c. TESTING FOR INTERACTION IN FUNCTIONAL VARYING-COEFFICIENT MODELS

Merve Yasemin Tekbudak*, *North Carolina State University*

Arnab Maity, *North Carolina State University*

The varying-coefficient model (VCM) is a useful alternative to the additive models for determining the relationship between the covariates and a response, while accounting for interactions. In this article, we consider a functional framework for varying-coefficient models, where a scalar response is regressed on a scalar covariate and a functional covariate, taking the interaction between these covariates into account. Our primary interest is to test for the nullity of the effect of the interaction. We develop the likelihood ratio and the generalized F testing procedures in this framework. Performance of these testing procedures are investigated in a simulation study. We further demonstrate our testing procedures on the Tecator data set.

✉ mytekbud@ncsu.edu

I0d. COMPARISON OF STATISTICAL METHODS TO CALCULATE A TEMPORAL BINDING WINDOW

John Bassler*, *West Virginia University and University of Alabama at Birmingham*

Sijin Wen, *West Virginia University*

Paula Webster, *West Virginia University*

James Lewis, *West Virginia University*

A study was conducted to identify differences in temporal multisensory integration between individuals with autism spectrum disorder (ASD) and neurotypical participants. Participants performed a task to gauge synchrony of auditory and visual stimuli presented at varying asynchronies. The perceived simultaneity of temporal synchrony was assessed by percentage of button presses for each time-point revealing a Gaussian like distribution used to derive a temporal binding window (TBW). The established method of TBW estimation calculates the area of the Gaussian curve below the 75% time within the maximum baseline value. This excluded participants whose TBWs were not estimable as their modeled curves were not within the 75% of baseline criterion. Our proposed method determines the TBW corresponding to the intersection of a vertical line where 75% of the curve area is contained and intersects the curve. TBW values within participants resulting from both methods were similar for participants for whom both methods were reliable; however, our new method enables the calculation of TBWs for participants with more atypical integration, enabling inclusion in analysis.

✉ johnrbassler@gmail.com

I0e. RANK BASED GROUP VARIABLE SELECTION FOR FUNCTIONAL REGRESSION MODEL

Jieun Park*, *Auburn University*

Ash Abebe, *Auburn University*

Nedret Billor, *Auburn University*

We propose a robust rank based variable selection method for a functional linear regression model with multiple explanatory functions and a scalar response. The procedure extends rank based group variable selection to functional variable selection and the proposed estimator is robust in the presence of outliers in predictor function space and response. The performance of the proposed robust method and their robustness are demonstrated with an extensive simulation study and real data examples.

✉ jzp0037@auburn.edu

10f. FUNCTIONAL DATA ANALYSES OF GAIT DATA MEASURED USING IN-SHOE SENSORS

Jihui Lee*, *Columbia University*

Gen Li, *Columbia University*

William F. Christensen, *Brigham Young University*

Gavin Collins, *Brigham Young University*

Matthew Seeley, *Brigham Young University*

Jeff Goldsmith, *Columbia University*

In gait studies, continuous measurement of force exerted by the ground on a body, or ground reaction force (GRF), provides valuable insights into biomechanics, locomotion, and possible presence of pathology. However, GRF requires a costly in-lab measurement obtained with sophisticated equipment. Recently, in-shoe sensors have been pursued as a relatively inexpensive alternative. In this study, we explore the properties of continuous in-shoe sensor recordings using functional data analysis approach. Our study is based on measurements of three healthy subjects, with more than 300 stances per subject. Sensor data show both phase and amplitude variability; we separate these sources via curve registration. We examine the correlation of phase shifts within a stance to evaluate the pattern of phase variability shared across sensors. We also compare the phase variability to detect potential similarities in phase shifts across subjects. Using registered curves, we explore possible associations in amplitude variability between in-shoe sensor recording and GRF measurement to evaluate the in-shoe sensor recording as a possible alternative to acquisition of GRF measurement.

✉ jl4201@cumc.columbia.edu

10g. USING DATA FROM MULTIPLE STUDIES TO DEVELOP A CHILD GROWTH CORRELATION MATRIX

Luo Xiao*, *North Carolina State University*

Craig Anderson, *University of Glasgow*

William Checkley, *Johns Hopkins University*

In many countries, the monitoring of child growth does not occur in a regular manner, and instead we may have to rely on sporadic observations which are subject to substantial measurement error. In these countries, it can be difficult to identify patterns of poor growth, and faltering children may miss out on essential health interventions. The contribution of this paper is to provide a framework for pooling together multiple datasets, thus allowing us to overcome the issue of sparse data and provide improved estimates of growth. We use data from multiple longitudinal growth studies to construct a common correlation matrix which can be used in estimation and prediction of child growth. The methodology utilizes statistical methods including functional data analysis, meta-analysis and smoothing. We illustrate the methodology using data from 16 child growth studies from the Bill and Melinda Gates Foundation's Healthy Birth Growth and Development knowledge integration (HBGDki) project.

✉ lxiao5@ncsu.edu

II. POSTERS: HIGH DIMENSIONAL DATA AND COMPUTATIONAL METHODS

IIa. SurVBoost: AN R PACKAGE FOR HIGH-DIMENSIONAL VARIABLE SELECTION IN THE STRATIFIED PROPORTIONAL HAZARDS MODEL VIA GRADIENT BOOSTING

Emily L. Morris*, *University of Michigan*

Zhi He, *University of Michigan*

Yanming Li, *University of Michigan*

Yi Li, *University of Michigan*

Jian Kang, *University of Michigan*

High-dimensional variable selection in the proportional hazards (PH) model has many successful applications in different areas. In practice, data may involve confounders that do not satisfy the PH assumption, in which case the stratified proportional hazards (SPH) model can be adopted to control the confounding effects by stratification of the confounder variable. However, there is lack of computationally efficient statistical software for high-dimensional variable selection in the SPH model. In this work, an R package, SurvBoost, is developed to implement the gradient boosting algorithm for fitting the SPH model with high-dimensional covariates and other confounders. Extensive simulation studies demonstrate that in many scenarios SurvBoost can achieve a better selection accuracy and reduce computational time substantially compared to the existing R package that implements boosting algorithms without stratifications. The proposed R package is also illustrated by an analysis of the gene expression data with survival outcome in The Cancer Genome Atlas (TCGA) study. In addition, a detailed hands-on tutorial for SurvBoost is provided.

✉ emorrisl@umich.edu

IIb. RENEWABLE ESTIMATION AND INFERENCE IN GENERALIZED LINEAR MODEL WITH STREAMING DATASETS

Lan Luo*, *University of Michigan*

Peter X.K. Song, *University of Michigan*

In this paper, we present an incremental Newton-Raphson learning algorithm to analyze streaming datasets in generalized linear models. Our proposed method is developed within a new framework of renewable estimation, in which the estimates can be renewed with current data and summary statistics of historical data, but with no use of any historical data themselves. In the implementation, we design a new data flow, named as the Rho architecture to accommodate the storage of summary statistics in the ancillary layer, as well as to communicate with the speed layer to facilitate sequential updating. We prove both estimation consistency and asymptotic efficiency of the renewable estimator, and proposed sequential updating for Wald test statistics and conduct inferences for model parameters. We illustrate our methods by various numerical examples from both simulation experiments and real-world analysis.

✉ luolsph@umich.edu

IIc. CONFIDENCE INTERVAL FOR THE RATIO OF TWO STANDARD DEVIATIONS OF NORMAL DISTRIBUTIONS WITH KNOWN COEFFICIENTS OF VARIATION

Wararit Panichkitkosolkul*, *Thammasat University, Thailand*

This paper proposes a confidence interval for the ratio of two standard deviations of normal distributions with known coefficients of variation. The phenomenon in which the coefficients of variation are known occurs in many fields such as agriculture, biology, and environmental and physical sciences. The proposed confidence interval is based on the approximate confidence interval for the ratio of normal means with a known coefficient of variation. A Monte Carlo

simulation study was conducted to compare the performance of the proposed confidence interval with the standard confidence interval based on the F-statistic. Simulation results showed that the proposed confidence interval performs much better than the standard confidence interval in terms of expected length. Moreover, the performance of both confidence intervals is illustrated through a real data example.

✉ wararit_tu@hotmail.com

IId. CLUSTERING MATRIX VARIATE DATA USING FINITE MIXTURE MODELS WITH COMPONENT-WISE REGULARIZATION

Peter A. Tait*, *McMaster University*

Paul D. McNicholas, *McMaster University*

Matrix variate distributions present an innate way to model random matrices. Realizations of random matrices are created by concurrently observing variables in different locations or at different time points. We use a finite mixture model composed of matrix variate normal densities to cluster matrix variate data that was generated by accelerometers worn by children in a clinical study. Their acceleration along the three planes of motion over the course of seven days, forms their matrix variate data. We use the resulting clusters to verify existing group membership labels derived from a test of motor-skills proficiency used to assess the children's locomotion.

✉ taitpa@mcmaster.ca

IIe. JACKKNIFE EMPIRICAL LIKELIHOOD INFERENCE FOR THE MEAN DIFFERENCE OF TWO ZERO INFLATED SKEWED POPULATIONS

Faysal I. Satter*, *Georgia State University*

Yichuan Zhao, *Georgia State University*

In finding a confidence interval for the mean difference of two populations, we may encounter the problem of having low coverage probability when there are many zeros in the data and the non-zero values are highly positively skewed. Because of the violation of the normality assumption, parametric methods are inefficient in such cases. One of the good alternative methods is the nonparametric empirical likelihood method. In this paper, jackknife empirical likelihood (JEL) and adjusted jackknife empirical likelihood (AJEL) methods are proposed to construct a nonparametric confidence interval for the mean difference of two zero inflated skewed populations. The confidence intervals by these two methods are compared with the confidence intervals by normal approximation method and empirical likelihood (EL) method proposed by Zhou & Zhou (2005). Simulation studies are carried out to assess the new methods. A real-life data set is also used as an illustration of the proposed methods.

✉ fsatter1@student.gsu.edu

IIf. APPLICATION OF NOVEL STATISTICAL METHOD FOR BODY POSTURE RECOGNITION USING WRIST-WORN ACCELEROMETER TO ASSESS DAILY STANDING PATTERNS IN HIV-INFECTED PATIENTS

Marcin Straczkiewicz*, *Indiana University*

Christopher Sorensen, *Washington University School of Medicine, St. Louis*

Beau Ances, *Washington University School of Medicine, St. Louis*

Jaroslaw Harezlak, *Indiana University*

Thanks to novel applications of statistical and signal processing techniques for accelerometry data we may now obtain more reliable estimates of physical activity in a free-living environment. We developed a method (SiSta – Sitting and Standing body posture recognition algorithm) to quantify standing time based on a tri-axial wrist-worn accelerometer. The key idea leverages the observation that hands are pointed down during standing and pointed mostly horizontally while sitting. In addition, standing activities commonly manifest in high sub-second variation. The

algorithm utilizes data obtained from one axis reflecting the spatial position of a sensor. We calculate the median and median standard deviation for the chosen axis in sliding time windows. The distinction between two states is obtained using the majority voting between indications provided by multiple logistic regression classifiers built using the aforementioned metrics in various time windows. The method is applied to data collected in a cohort of HIV-infected individuals who wore the devices for one week at baseline and at a 3-month post-randomization visit.

✉ mstraczk@iu.edu

IIg. SAFE-CLUSTERING: SINGLE-CELL AGGREGATED (FROM ENSEMBLE) CLUSTERING FOR SINGLE-CELL RNA-Seq DATA

Yuchen Yang*, *University of North Carolina, Chapel Hill*

Ruth Huh, *University of North Carolina, Chapel Hill*

Houston Culpepper, *University of North Carolina, Chapel Hill*

Yun Li, *University of North Carolina, Chapel Hill*

Several methods have been developed for cell type identifying from single-cell RNA-seq (scRNA-seq) data. However, it is difficult to select an optimal method as no individual clustering method is the obvious winner across various datasets. Here, we present SAFE-clustering, Single-cell Aggregated (From Ensemble) clustering, an accurate and robust method for clustering scRNA-seq data. SAFE-clustering takes as input results from multiple clustering methods to one consensus clustering. Single cells are first clustered using four popular methods, SC3, CIDR, Seurat and t-SNE + k-means; individual solutions are then ensembled using three hypergraph-based partitioning algorithms, hypergraph partitioning algorithm (HGPA), meta-cluster algorithm (MCLA) and cluster-based similarity partitioning algorithm (CSPA). We performed evaluation across 14 datasets with cell number ranging from 49 to 32,695. In our evaluations, SAFE-clustering generates high-quality clustering across various datasets, in terms of both cluster number and cluster assignment. Moreover, SAFE-clustering is computationally efficient. For example, it takes less than 10 seconds to cluster 28,733 cells by MCLA algorithm.

✉ yyuchen@email.unc.edu

IIh. THE N-LEAP METHOD FOR STOCHASTIC SIMULATION OF COUPLED CHEMICAL REACTIONS

Yuting Xu*, *Merck & Co.*

Yueheng Lan, *Beijing University of Posts and Telecommunications*

Numerical simulation of the time evolution of a spatially homogeneous chemical system is always of great interest. Gillespie first developed the exact stochastic simulation algorithm (SSA), which is accurate but time-consuming. Recently, many approximate schemes of the SSA are proposed to speed up simulation. Presented here is the N-leap method, which guarantees the validity of the leap condition and at the same time keeps the efficiency. In many cases, N-leap has better performance than the widely-used τ-leap method. The details of the N-leap method are described and several examples are presented to show its validity.

✉ xuyuting1990@gmail.com

IIi. THE EFFICIENCY OF RANKING COUNT DATA WITH EXCESS ZEROS

Deborah A. Kanda*, *Georgia Southern University*

Jingjing Yin, *Georgia Southern University*

Data from public health studies often include count endpoints that exhibit excess zeros and depending on the study objectives, hurdle or zero-inflated models are used to model such data. In this study, we propose to apply a sampling scheme that is based on ranking, which significantly reduces the sample size and thus study cost for count data with excess zero. The appeal of ranked set sampling is its

ability to give more precise estimation than simple random sampling as ranked set samples (RSS) are more likely to span the full range of the population. Intensive simulations are conducted to compare the proposed sampling method using RSS with simple random samples (SRS), comparing the mean squared error (MSE), bias, variance, and power of the RSS with the SRS under various data generating scenarios. We also illustrate the merits of RSS on a real data set with excess zeros using data from the National Medical Expenditure Survey on demand for medical care. Results from data analysis and simulation study coincide and show the RSS outperforming the SRS in all cases, with the RSS showing smaller variances and MSE compared to the SRS.

✉ dk01636@georgiasouthern.edu

IIj. NEIGHBORHOOD SELECTION WITH APPLICATION TO SOCIAL NETWORKS

Nana Wang*, *University of California, Davis*

Wolfgang Polonik, *University of California, Davis*

The topic of this presentation is modeling and analyzing dependence in social networks, which has not received much attention in the literature so far. We propose a latent variable block model that allows the analysis of dependence between blocks via the analysis of a latent graphical model. Our approach is based on the idea underlying the neighborhood selection scheme put forward by Meinshausen and Bühlmann. However, because of the latent nature of our model, estimates have to be used in lieu of the unobserved variable. This leads to a novel analysis of graphical models under uncertainty, in the spirit of Rosenbaum and Tsybakov (2010), or Belloni, Rosenbaum and Tsybakov (2016). Lasso-based selectors, and a class of Dantzig-type selectors are studied.

✉ nnawang@ucdavis.edu

IIk. SIMULATION STUDY AND APPLICATIONS OF THE BURR XII WEIBULL DISTRIBUTION

Sarah Ayoku*, *Georgia Southern University*

Broderick Oluyede, *Georgia Southern University*

In this poster, a simulation study is conducted to examine the performance of the Maximum Likelihood Estimates (MLE) of the Burr-XII Weibull (BXIIW) distribution using different sample sizes and parameter values. We examine the bias, mean square error and width of the confidence interval for each parameter. Applications of the distribution to real data sets in order to illustrate the flexibility of the BXIIW (referred to as BW) distribution and its sub-models for data modeling as well as comparison with the non-nested gamma-Dagum (GD) are presented.

✉ sa03348@georgiasouthern.edu

I2. POSTERS: PREDICTION, DIAGNOSTICS, AND RISK FACTOR IDENTIFICATION

I2a. INFECTIOUS DISEASE DETECTION USING SPECIMEN POOLING WITH MULTIPLEX ASSAYS WHEN RISK-FACTOR INFORMATION IS PRESENT

Christopher R. Bilder*, *University of Nebraska-Lincoln*

Joshua M. Tebbs, *University of South Carolina*

Christopher S. McMahan, *Clemson University*

High-volume testing of clinical specimens for infectious diseases is performed by laboratories across the world. To make testing loads manageable, laboratories frequently employ the use of group testing (tests performed on pools of specimens) with multiplex assays (multiple-disease tests). In our presentation, we propose incorporating individual risk-factor information, such as exposure history and clinical observations, into this testing process. We show that significant gains in testing efficiency can be obtained

in comparison to current testing procedures. Our application focus is on the Aptima Combo 2 Assay that is used by laboratories for chlamydia and gonorrhea testing.

✉ chris@chrisbilder.com

I2b. A GROUP TESTING DESIGN THAT CAN ACHIEVE MORE BY DOING LESS

Dewei Wang*, *University of South Carolina*

This study focuses on group testing experiments where the goal is to estimate covariate effects through logistic regression. When the assay is perfect, collecting data by testing individual specimens separately is the most effective way in terms of the resulting estimators (of the regression coefficients) are the most efficient. However, this is not true when the assay is not perfect. It actually can be shown that when the prevalence is low, group testing can yield more efficient estimators than individual testing while using a much less number of tests. In order to maximize this counterintuitive advantage, a new two-step informative group testing design is proposed. Theoretical conditions for using the new design are provided. The performance of the new design, when comparing to individual testing, is illustrated through simulation and a chlamydia data set.

✉ deweiwang@stat.sc.edu

I2c. STATISTICAL MODELS FOR PREDICTING KNEE OSTEOARTHRITIS ENDPOINTS: DATA FROM THE OSTEOARTHRITIS INITIATIVE

Robin M. Dunn*, *Carnegie Mellon University*

Joel Greenhouse, *Carnegie Mellon University*

Peter Mesenbrink, *Novartis Pharmaceuticals Corporation*

David James, *Novartis Pharmaceuticals Corporation*

David Ohlssen, *Novartis Pharmaceuticals Corporation*

Osteoarthritis (OA) is a degenerative joint disease for which there is currently no approved treatment to stop or reverse disease progression. Knee OA is the most common form of OA. The Osteoarthritis Initiative (OAI) conducted a ten-year longitudinal study to examine the onset and progression of knee OA. Once disease progression is sufficiently severe in a knee, often the only option is to replace the knee. To predict which individuals in the OAI study will undergo knee replacement, this work compares logistic regression, logistic regression with a Lasso penalty, classification random forests, and regression random forests. Using 10-fold stratified cross-validation to estimate model performance, it is discovered that knee replacements can be predicted with high degrees of sensitivity and specificity. As an alternative endpoint, end-stage knee OA is also examined, using a definition from Driban et al. (2016) that takes into account symptomatic and radiographic progression. Models for the time to event of end-stage knee OA are constructed, incorporating methods such as dimension reduction and imputation of missing data.

✉ dunnr@cmu.edu

I2d. OPEN SOURCE MACHINE LEARNING ALGORITHMS FOR PREDICTION OF OPTIMAL CANCER DRUG THERAPY

Cai Huang*, *Georgia Institute of Technology*

Precision medicine is a rapidly growing area of modern medical science and open source machine learning codes promise to be a critical component for the successful development of standardized and automated analysis of patient data. One important goal of precision cancer medicine is the accurate prediction of optimal drug therapies from the genomic profiles of individual patient tumors. We introduce here an open source software platform that employs a highly versatile SVM algorithm combined with recursive feature elimination (RFE) approach. Drug-specific models were built using gene expression and drug response data from the NCI-60. The models are highly accurate in predicting the drug responsiveness of a variety of cancer cell lines including those comprising the recent NCI-DREAM Challenge. We demonstrate that predictive accuracy is optimized when

the learning dataset utilizes all probe-set expression values from a diversity of cancer cell types without pre-filtering for genes generally considered to be "drivers" of cancer onset/progression.

✉ chuang95@gatech.edu

I2e. COPAS-LIKE SELECTION MODEL TO CORRECT PUBLICATION BIAS IN SYSTEMATIC REVIEW OF DIAGNOSTIC TEST STUDIES

Jin Piao*, *University of Southern California*

Yulun Liu, *University of Pennsylvania*

Yong Chen, *University of Pennsylvania*

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

The accuracy of a diagnostic test, which is often quantified by a pair of measures such as sensitivity and specificity, is critical for medical decision making. Separate studies of an investigational diagnostic test can be combined through meta-analysis; however, such an analysis can be threatened by publication bias. To the best of our knowledge, there is no existing method that accounts for publication bias in the meta-analysis of diagnostic tests involving bivariate outcomes. In this paper, we extend the Copas selection model from univariate outcomes to bivariate outcomes for the correction of publication bias when the probability of a study being published can depend on its sensitivity, specificity, and the associated standard errors. We develop an expectation-maximization algorithm for the maximum likelihood estimation under the proposed selection model. We investigate the finite sample performance of the proposed method through simulation studies and illustrate the method by assessing a meta-analysis of 17 published studies of a rapid diagnostic test for influenza.

✉ jpiao@usc.edu

I2f. USING BIOSTATISTICS TO IMPROVE HEALTH OUTCOMES ON THE LAST MILE OF A LEARNING HEALTHCARE SYSTEM

Daniel W. Byrne*, *Vanderbilt University*

Henry J. Domenico, *Vanderbilt University*

Li Wang, *Vanderbilt University*

Biostatisticians routinely perform important work in analyzing healthcare data and creating predictive models. Although this work often leads to publications in medical research journals, the findings seldom lead to improved, sustainable health outcomes. When findings are applied, average time lag from discovery to practice is 17 years. To overcome these problems, we created a Learning Healthcare System Platform. Using this infrastructure, we successfully completed several large randomized pragmatic trials of: 1) daily chlorhexidine bathing of ICU patients (n=9,340); 2) a comparison of normal saline vs. balanced fluids (n=29,149); 3) the effectiveness of a post-discharge phone call on hospital readmissions (n=2,738). We have also tested whether real-time predictive models incorporated into the electronic health record can be used to focus prevention and improve outcomes in randomized controlled trials. We have demonstrated that it is possible to implement rigorous, randomized controlled trials in a hospital setting. As well as describing the infrastructure and processes, we discuss the initial concerns, previous obstacles, and the remaining challenges.

✉ daniel.byrne@vanderbilt.edu

I2g. USING FRAILTY MODELS TO IMPROVE BREAST CANCER RISK PREDICTION

Theodore J. Huang*, *Harvard School of Public Health and Dana-Farber Cancer Institute*

Danielle Braun, *Harvard School of Public Health and Dana-Farber Cancer Institute*

Malka Gorfine, *Tel Aviv University*

Li Hsu, *Fred Hutchinson Cancer Research Center*

Giovanni Parmigiani, *Harvard School of Public Health and Dana-Farber Cancer Institute*

There are numerous statistical models used to identify individuals at high risk of cancer due to inherited mutations. One such model is BRCAPRO, a Mendelian risk prediction model that predicts future risk of breast and ovarian cancers by estimating the probabilities of germline deleterious mutations in the BRCA1 and BRCA2 genes, which is the focus of this work. BRCAPRO makes these predictions by incorporating family history, estimated disease penetrance's and mutation prevalence's, and other factors such as race and prophylactic surgeries. However, one limitation of the BRCAPRO model is its inability to account for the heterogeneity of risk across families due to sources such as environmental or unobserved genetic risk factors. We aim to improve breast cancer risk prediction in the model by incorporating a frailty model that contains a family-specific frailty vector to account for this heterogeneity. We apply our proposed model to data from the Cancer Genetics Network and show improvements in model calibration and discrimination.

✉ thuang01@g.harvard.edu

I2h. AN ENSEMBLE MODELING APPROACH FOR ESTIMATING GLOBAL BURDEN OF INFLUENZA-ASSOCIATED HOSPITALIZATIONS

Mingrui Liang*, *Emory University*

Howard Chang, *Emory University*

Guandong Yang, *Emory University*

Katherine Roguski, *Centers for Disease Control and Prevention*

Jeremy Reich, *Centers for Disease Control and Prevention*

Danielle Iuliano, *Centers for Disease Control and Prevention*

Neha Patel, *Emory University*

Vanessa Cozza, *World Health Organization*

Julia Fitzner, *World Health Organization*

One major analytic challenge in assessing the global burden of disease is the need to extrapolate available burden estimates from a small number of countries to other countries without the necessary data to calculate these estimates. Motivated by a recent project to estimate global influenza-associated hospitalizations, we developed an ensemble approach to perform extrapolation. First, for countries with annual rate estimates, an average rate is estimated for each country using a Bayesian hierarchical model. Extrapolated rates for countries without data are obtained as a mixture of posterior distributions of average rates from countries with data. Mixture weights are defined by a distance measure of principle components derived from a large number of country-specific covariates related to population health. The posterior distribution of a country is given more weights if its principle component values are similar to those of the country to be extrapolated. Using cross-validation experiments, we evaluate the use of different number of principle components and distance metrics to determine optimal weights based on 3 different measurements.

✉ mliang4@emory.edu

I2i. NET BENEFIT CURVES: A MODEL PERFORMANCE MEASURE FOR EXAMINING CLINICAL USEFULNESS

Anwesha Mukherjee*, *Florida State University*

Daniel L. McGee, *Florida State University*

ROC curves are generally used to evaluate accuracy of prediction models. The objective of this study is to find measures which not only incorporate statistical but also clinical consequences of a model. Depending on the disease and study population, the misclassification costs of false positives and false negatives vary. Decision Curve Analysis (DCA) takes this cost into account, by using a

threshold (probability above which a patient opts for treatment). Using the DCA technique, Net Benefit Curve is built by plotting Net Benefit, a function of expected benefit and expected harm of a model, against threshold. Threshold ranges relevant to the disease and study population is used in the net benefit plot for optimum results. A summary measure is constructed to find which model yields highest net benefit. The most intuitive approach is to calculate area under the curve. It needs to be examined if use of weights creates a better summary measure. Several datasets are used to compute the measures, Area under ROC, Area under Net Benefit Curve and Weighted Area under Net Benefit Curve. Results from these analyses reveal significant variability among studies.

✉ a.mukherjee90@stat.fsu.edu

I3. POSTERS: CLINICAL TRIALS AND BIOPHARMACEUTICAL RESEARCH METHODS

I3a. ADAPTIVE BAYESIAN PHASE I CLINICAL TRIAL DESIGN FOR ESTIMATION OF MAXIMUM TOLERATED DOSES OF TWO DRUGS WHILE FULLY UTILIZING ALL TOXICITY INFORMATION

Yuzi Zhang*, *Emory University*

Michael Kutner, *Emory University*

Jeanne Kowalski, *Emory University*

Zhengjia Chen, *Emory University*

We develop a Bayesian adaptive Phase I clinical trial design entitled Escalation with Overdoing Control using Normalized Equivalent Toxicity Score for estimating maximum tolerated dose (MTD) contour of two drug Combination (EWOC-NETS-COM). The normalized equivalent toxicity score (NETS) as the primary endpoint of clinical trial is assumed to follow quasi-Bernoulli distribution and treated as quasi-continuous random variable in the logistic likelihood function. Four parameters with explicit clinical meanings are re-parameterized to describe the association between NETS and dose levels of two drugs. Non-informative priors are used and Markov chain Monte Carlo is employed to update the posteriors of the 4 parameters. Extensive simulations are conducted to evaluate the safety, therapeutic effect, and trial efficiency of EWOC-NETS-COM under different scenarios, using the EWOC as reference. The results demonstrate that EWOC-NETS-COM not only estimates MTD contour of multiple drugs, but also provides better therapeutic effect by reducing the probability of treating patients at underdose and fully utilizes all toxicity information to improve trial efficiency.

✉ yuzi.zhang@emory.edu

I3b. ANALYSES OF LONGITUDINAL CLINICAL DATA WITH TIME-VARYING COVARIATES IN LARGE AND LONG-TERM TRIALS

Qianyi Zhang*, *Eli Lilly and Company*

Rong Liu, *Eli Lilly and Company*

For clinical trials with a time-to-rare event outcome, large sample sizes and long-term data collection are usually required. Patients' characteristics are collected at multiple visits until the censoring time or the occurrence of the key clinical event. However, the treatment effect after randomization is often analyzed without considering changes in patients' profiles over time. When such trials fail to meet the primary endpoint to show superiority of the study treatment in survival, it is of strong interest to investigate what can really predict the survival. We proposed to use a discrete-time survival tree (Bou-Hamad, 2011) with time-varying covariates to identify important risk factors from all data collected at baseline, during the study and at event onset time. This data-driven method allows correlated covariates and does not impose any assumptions for the relationship between outcome and covariates. We have used this approach to identify risk factors associated with cardiovascular events in Type 2 diabetic patients (Strojek, 2016).

✉ liurongrl@lilly.com

I3c. TWO-STAGE DESIGN CONSIDERING SUPERIORITY AND NON-INFERIORITY TESTS IN THREE-ARM CLINICAL TRIALS

Yoshikazu Ajisawa*, *Tokyo University of Science*

Shogo Nomura, *National Cancer Center, Japan*

Takashi Sozu, *Tokyo University of Science*

We assume a three-arm clinical trial with two treatment groups (T1 and T2) and a control group (C), and consider the superiority of T1 or T2 to C and the non-inferiority of T2 to T1. In such a clinical trial, the power for demonstrating the non-inferiority tends to be insufficient using the number of participants needed for demonstrating the superiority. We propose a two-stage design to enroll the required number of participants efficiently, assuming that the efficacy of T1 is greater than that of T2. In the first stage, the required number of participants are enrolled for demonstrating the superiority. If superiority is demonstrated, we will stop the enrollment of participants in the control group and proceed to the second stage. In the second stage, the required number of participants are enrolled for demonstrating the non-inferiority, where the data from the T1 and T2 at the first stage is used for the analysis. We show the numerical examples of the probability of demonstrating the superiority and/or non-inferiority and the expected number of participants under the proposed two-stage design and a conventional design.

✉ 4416601@ed.tus.ac.jp

I3d. OPTIMAL SAMPLE SIZE FOR CLUSTER RANDOMIZED TRIALS BASED ON SIMULATION AND RANDOMIZATION DISTRIBUTION

Ruoshui Zhai*, *Brown University*

Roee Gutman, *Brown University*

In cluster randomization trials (CRTs), groups are randomized to treatments rather than individuals to minimize experimental contamination. However, with a common intention to make individual-level inference, CRTs could be less efficient than randomizing individuals directly since members in the same cluster could be correlated. With such unique characteristics, sample size calculation for CRTs calls for special attention. Many papers propose formulas relying on asymptotic normality of the test statistic under the randomization and sampling distributions. However, commonly in CRTs, only a small number of groups are actually randomized, and thus the asymptotic normality may not be appropriate. Moreover, when the outcomes are recorded as discrete values the normality assumption is even harder to justify. We present an approach to calculate the required sample size for CRTs by approximating the distribution of any test statistic with simulation. This approach is general, non-parametric, and does not limit investigators to specific types of outcomes or test statistics. It can also incorporate the ratio of budgets in a two-armed trial and take existing data as a baseline.

✉ ruoshui_zhai@brown.edu

I3e. QUANTAL RESPONSE DATA ANALYSIS WITH COVARIATES

Lili Tong*, *University of South Carolina*

Edsel Pena, *University of South Carolina*

This is a problem dealing with quantal response finding in the context of having covariates and also when a portion of the semi-parametric model could belong to several model classes. In the previous, researchers did not consider covariates, but having covariates is a natural extension which will be more useful in practice. This project would closely look at the covariate cases and decide the case of several model classes. Moreover, there is an inherent ordering of adverse event probabilities, hence isotonic type regression is an important component of the solution. An application of the methods can be utilized in toxicology studies.

✉ tonglili91@hotmail.com

I3f. MODIFIED WALD TESTS FOR REFERENCE SCALED EQUIVALENCE ASSESSMENT OF ANALYTICAL BIOSIMILARITY

Yu-Ting Weng*, *U.S. Food and Drug Administration*

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

Meiyu Shen, *U.S. Food and Drug Administration*

Chao Wang, *U.S. Food and Drug Administration*

For the reference scaled equivalence hypothesis, Chen et al. (2017) proposed to use Wald test with Constrained Maximum Likelihood Estimate (CMLE) of the standard error to improve the efficiency when the numbers of lots for both test and reference products are small and variances are unequal. However, by using the Wald test with CMLE standard error (Chen et al., 2017), simulations show that the type I error rate is below the nominal significance level. Weng et al. (2017) proposed the Modified Wald test with CMLE standard error by replacing the maximum likelihood estimate of reference standard deviation with the sample estimate (MW-CMLE), resulting in further improvement of type I error rate and power over the tests proposed in Chen et al. (2017). In this presentation, we further compare the proposed method to exact-test-based method (Dong et al., 2017a) and Generalized Pivotal Quantity (GPQ) method (Weerahandi, 1993) with equal or unequal variance ratios or equal or unequal numbers of lots for both products. The simulations show that the proposed MW-CMLE method outperforms other two methods in terms of type I error rate control and power improvement.

✉ Yu-Ting.Weng@fda.hhs.gov

I3g. POWER AND TYPE I ERROR RATE ESTIMATION VIA SIMULATION FOR MULTISTATE ENDPOINTS

Ryan A. Peterson*, *University of Iowa*

Jennifer G. Rademacher, *Mayo Clinic*

Sumithra J. Mandrekar, *Mayo Clinic*

Terry M. Therneau, *Mayo Clinic*

Multistate modeling is a useful complement to standard survival methods that can more completely capture the patient trajectory over the course of a clinical trial. In some clinical trials, treatments affect a patient in different ways at multiple states, and detecting this treatment pattern is of high clinical importance. In such cases, the power to detect a difference in multistate endpoints depends substantially on study-specific multistate structures and the transition rates between possible states. Further, since more than one outcome exists for multistate models, we must carefully consider and correct for the issue of multiple comparisons. We advocate a guided simulation process to estimate the power and the type I error rate for clinical trials with multistate endpoints. We introduce the Multistate Simulation Designer shiny application: an open-sourced R project to guide clinical trialists and statisticians through the multistate simulation process. The Multistate Simulation Designer allows users to explore various, highly customizable multistate structures and simulate clinical trials in an efficient, reproducible, and user-friendly fashion.

✉ ryan-peterson@uiowa.edu

I3h. AN APPLICATION OF AUGMENTED BETA REGRESSION TECHNIQUES IN PHARMACOKINETIC-PHARMACODYNAMIC MODELING OF A BOUNDED OUTCOME IN PSORIASIS

James A. Rogers*, *Metrum Research Group*

Jonathan French, *Metrum Research Group*

Bojan Lalovic, *Eisai Co. Ltd.*

Population pharmacokinetic-pharmacodynamic (pop PK-PD) models are used broadly in drug development to characterize and anticipate the effects of various dosing regimens in a population of patients. In cases where pharmacodynamic endpoints are bounded, Pop PK-PD models utilizing a Beta or augmented Beta residual structure have obvious appeal, particularly given a desire for realistic model-based simulations that respect the constraints on the endpoint. Beta regression techniques have only recently

been introduced in this context, and effective model checking strategies remain unelaborated. In this work we describe an approach to model checking and model refinement that we applied to develop a pop PK-PD model for the Psoriasis Area and Severity Index (PASI), a scale bounded below by zero and bounded above at 72.

✉ jimr@metrumrg.com

I3i. BREAST CANCER MULTI-ARM CLINICAL TRIALS WITH MULTIPLE OBJECTIVES: A LITERATURE REVIEW OF MAJOR JOURNALS

Yu Miyauchi*, *Tokyo University of Science*

Shogo Nomura, *National Cancer Center, Japan*

Yoshikazu Ajisawa, *Tokyo University of Science*

Takashi Sozu, *Tokyo University of Science*

Breast cancer multi-arm trials attract interest, but tend to be complex because they generally have multiple study objectives. We conducted a literature review to elucidate the existing status of study designs and analysis methods in such trials. A search in PubMed databases identified 455 articles including the search words "breast" and "randomized (or randomised)" from 2010 to 2016. The targeted journals were the New England Journal of Medicine, the Lancet, the Lancet Oncology, and the Journal of Clinical Oncology. We collected the necessary information on study designs and analysis methods from the above four articles, corresponding study protocols and/or statistical analysis plans, if available, and four online registry systems (e.g., ClinicalTrials.gov). A total of 51 articles (44 trials) were selected. Many trials were multi-regional (59%), academic-sponsored (52%), large-scaled (more than 500 participants; 50%), and drug-evaluated (75%) studies. The shared-control design, which randomized one control and several experimental treatments and applied multiplicity adjustment methods, was frequently used in the three-armed trials (77%).

✉ 4417623@ed.tus.ac.jp

I4. POSTERS: IMAGING AND NEUROSCIENCE

I4a. IMAGE-ON-IMAGE REGRESSION: A SPATIAL BAYESIAN LATENT FACTOR MODEL FOR PREDICTING TASK-EVOKED BRAIN ACTIVITY USING TASK-FREE MRI

Cui Guo*, *University of Michigan*

Jian Kang, *University of Michigan*

Timothy D. Johnson, *University of Michigan*

Our brains have markedly different activity during task performance in almost all behavioral domains, attributed to possible factors, 1) differences in gross brain morphology and 2) different task strategy or cognitive processes. We assume individual differences in task-evoked brain activity are, to a great degree, inherent features of individual brain. Then, a research question of interest is whether task-free MRI can be used to predict several task-evoked activity maps in multiple behavioral domains. We propose an image-on-image regression model, which is also a spatial Bayesian latent factor regression model. The task-evoked maps and their spatial correlations are measured through a collection of basis functions. The low-dimensional representation of the basis parameters is obtained by placing a sparse latent factor model. Then we use a scalar-on-image regression model to link the latent factors with a few task-free maps selected by using Bayesian variable selection method. Our proposed model is applied to 98 subjects of the Human Connectome Project (HCP) database, who have contrast maps in 7 behavioral domains and 107 task-free MRI.

✉ cuiguo@umich.edu

I4b. FEATURES EXTRACTION IN BRAIN VIDEO VIA DEEP CONVOLUTIONAL AUTOENCODER

Jiahui Guan*, *University of California, Davis*

Feature extraction has long been important in many fields such as medical imaging and neuronscience. However, with more and more large and complex datasets being gathered, it is impossible to manually extract features without any prior knowledge, especially in the presence of noise and subjects' movement. In the past few years, deep learning, specifically autoencoder, has been used to discover representations of features. Compared to PCA, autoencoder requires minimal assumption. Nevertheless, in medical brain video where images in different time points vary slightly, traditional autoencoder may neglect those small changes. Driven by this significance, we propose a new autoencoder structure that can capture feature differences no matter huge or subtle in this paper. The general network architecture contains a contracting path followed by an expensive path. A concatenation with the correspondingly cropped feature map from the contracting path is used. Experiments are conducted on real-world medical brain video clips and the results demonstrate the effectiveness and efficiency of our proposed structure.

✉ jiaguan@ucdavis.edu

I4c. POWERFUL PERMUTATION TESTS FOR NEUROIMAGING USING VOXEL-WISE TRANSFORMATIONS

Simon N. Vandekar*, *University of Pennsylvania*

Theodore D. Satterthwaite, *University of Pennsylvania*

Adon Rosen, *University of Pennsylvania*

Rastko Ciric, *University of Pennsylvania*

David R. Roalf, *University of Pennsylvania*

Kosha Ruparel, *University of Pennsylvania*

Ruben C. Gur, *University of Pennsylvania*

Raquel E. Gur, *University of Pennsylvania*

Russell T. Shinohara, *University of Pennsylvania*

Typical statistical methods in neuroimaging result in hundreds of thousands of tests performed in an image, followed by the use of a multiple testing procedure (MTP) to control the family-wise error rate (FWER). Recent studies have demonstrated that widely used MTP procedures yield anticonservative FWERs. The permutation MTP is among few procedures shown to reliably control the number of false positives at the specified probability. Voxel-wise permutation tests work by randomly permuting the imaging data and using the distribution of the maximum value of the test statistic across all voxels in the image to compute adjusted p-values. While this procedure has intuitive appeal, anecdotally many investigators have noted it lacks power. We demonstrate that the procedure lacks power because neuroimaging data have voxels with heavy skew near the edge of the brain. These voxels cause the distribution of the maximum across the image to be heavily inflated. As a solution we apply the Yeo-Johnson transformation prior to permutation testing. The transformation yields a statistical image where all the voxels have approximately the same distribution and improves the power of the test.

✉ simonv@pennmedicine.upenn.edu

I4d. MULTIPLE TESTING BASED ON SEMI-PARAMETRIC HIERARCHICAL MIXTURE MODELS UNDER DEPENDENCY IN DISEASE-ASSOCIATION STUDIES WITH NEUROIMAGING DATA

Ryo Emoto*, *Nagoya University*

Atsushi Kawaguchi, *Saga University*

Hisako Yoshida, *Saga University*

Shigeyuki Matsui, *Nagoya University*

In multiple testing to associate neuroimaging data with disease status, it is important to consider the spatial structure in the neuroimaging data. However, the conventional multiple testing approaches to voxel-level inference often ignore the spatial dependency and induce substantial loss of efficiency. Recently a method using local index of significance based on a finite normal mixture model with hidden Markov random field structure is proposed to incorporate the spatial dependency. In this paper, we propose a method based on a hierarchical mixture model with a nonparametric prior distribution, rather than the mixture of normal distributions, to more flexibly estimate the underlying effect size distribution, possibly with a wide range of distributional forms. Our method also allows for flexible estimation of effect sizes for individual voxels. Simulation results demonstrate that our method can estimate the effect size of each voxel more accurately than the previously developed method based on finite normal mixture models, especially when the true effect size distribution largely departs from normal distributions.

✉ emoto.ryo@b.mbox.nagoya-u.ac.jp

I4e. STATISTICAL INFERENCE FOR THE FIRST PASSAGE TIME OF A DIFFUSION PROCESS OF NEURAL ACTIVITY

Bowen Yi*, *University of Pittsburgh*

Satish Iyengar, *University of Pittsburgh*

In this paper, we investigate the statistical inference problem for the first passage time of a Feller process. The motivation comes from an integrate-and-fire neuron model that includes an inhibitory reversal potential. This model also arises in mathematical finance, where it is known as the Cox-Ingersoll-Ross for the evolution of interest rates. The aim here is to construct a maximum likelihood estimator of the parameters given only the renewal process consisting of first passage time observations for an unknown constant boundary. We study the density functions of the first passage time through its Laplace transform, which is known to be a ratio of confluent hypergeometric functions. We show the identifiable parameters and verify the regularity conditions of a conditional version of maximum likelihood estimation. We then compute the density function along with its derivatives through an inversion formula and provide an interval estimator for the unknown parameters; we also present results of simulation studies and study actual neural spike trains.

✉ boy9@pitt.edu

I4f. GLOBAL PCA OF LOCAL MOMENTS WITH APPLICATIONS TO MRI SEGMENTATION

Jacob M. Maronge, University of Wisconsin, Madison*

John Muschelli, *Johns Hopkins Bloomberg School of Public Health*

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

We are interested in describing the information contained in local neighborhoods, and higher moments of local neighborhoods, of complex multimodal imaging techniques at the population level. This is problematic because of the size of medical imaging data. We propose a simple, computationally-efficient approach for representing the variation in multimodal images using the spatial information contained in all local neighborhoods across multiple subjects. This method achieves 3 goals: 1) decomposes the observed variability images at the population level; 2) describes and quantifies the main directions of variation; 3) uses these directions of variation to improve segmentation and studies of association with health outcomes. To achieve this, we efficiently decompose the observed variation in local neighborhood moments. In order to assess the quality of this method we show results using the 2015 Ischemic Stroke Lesion Segmentation (ISLES) Challenge.

✉ jmmaronge@gmail.com

I4g. A NON-PARAMETRIC, NETWORK-BASED TEST FOR GROUP DIFFERENCES IN MULTIVARIATE SUBJECT-LEVEL HISTOGRAMS

Jordan D. Dworkin*, *University of Pennsylvania*

Russell T. Shinohara, *University of Pennsylvania*

Using multi-modal MRI, diffuse disease processes in the brain can result in differences in the structure of subjects' multivariate voxel intensity histograms. Yet current methods for quantifying group differences in image characteristics typically rely on summary statistics, which can obscure distributional differences and ignore variations in histogram sizes, or region-specific analyses, which can erase non-localized processes. We propose a non-parametric test based on graph-theoretic principles that utilizes subject-level histograms, accounts for differences in histogram size, and allows for variability in the location of the disease process. This method can be carried out either globally or locally, with local testing allowing for the visualization of voxel types that contribute to group differences. Simulations reveal that the proposed method controls type I error and has higher power than tests comparing group-level histograms. This method represents an alternative to the use of summary statistics or voxel-level analysis when investigating diffuse multivariate processes in the brain, and its performance demonstrates the potential of network-based testing.

✉ jdwor@pennmedicine.upenn.edu

I5. POSTERS: METHODS FOR CATEGORICAL AND ORDINAL DATA

I5a. A SMOOTH NONPARAMETRIC APPROACH TO DETERMINING CUT-POINTS OF A CONTINUOUS SCALE

Zhiping Qiu*, *Emory University*

Limin Peng, *Emory University*

Amita Manatunga, *Emory University*

Ying Guo, *Emory University*

The problem of determining cut-points of a continuous scale according to an establish categorical scale is often encountered in practice for the purposes such as making diagnosis or treatment recommendation, determining study eligibility, or facilitating interpretations. While simple non-parametric estimators and the associated theory are derived in Peng et al. (2016), the implementation of their method can be computationally intensive when more than a few cut-points need to be determined. In this work, we propose a smoothing-based modification of Peng et al. (2016)'s estimation procedure that can substantially improve the computational speed and asymptotic convergence. Moreover, we develop plug-in type variance estimation for the new nonparametric estimator, which further facilitates the computation. Extensive simulation studies confirm our theoretical results and demonstrate the computational benefits of the new method. The practical utility of our proposal is illustrated by an application to a mental health study.

✉ zqiu8@emory.edu

I5b. TESTING HOMOGENEITY OF DIFFERENCE OF TWO PROPORTIONS FOR STRATIFIED CORRELATED PAIRED BINARY DATA

Xi Shen*, *State University of New York at Buffalo*

Changxing Ma, *State University of New York at Buffalo*

In ophthalmologic or otolaryngologic study, each subject may contribute paired organs measurements to the analysis. A number of statistical methods have been proposed on bilateral correlated data. In practice, it is important to detect confounding effect by treatment interaction, since ignoring confounding effect may lead to unreliable conclusion. Therefore, stratified data analysis can be considered to adjust the effect of confounder on statistical inference. In this article, we investigate and derive three test procedures for testing homogeneity of difference of two proportions for stratified correlated paired binary data in the basis of

equal correlation model assumption. The performance of proposed test procedures is examined through Monte Carlo simulation. The simulation results show that the score test is usually robust on type I error control with high power, and therefore is recommended among the three methods. One example from otolaryngologic study is given to illustrate the three test procedures.

✉ xishen@buffalo.edu

I5c. ASSESS TREATMENT EFFECTS FOR MULTIPLE GROUPS FOR ORDINAL OUTCOME WHEN CONFOUNDING EXISTS

Soutik Ghosal*, *University of Louisville*

Maiying Kong, *University of Louisville*

Ordinal data are very common in clinical fields, and the very basic research question is to assess the treatment effects from two or more treatments. Randomized Controlled Trials (RCT) are considered as a gold standard approach to estimate the treatment effect. Many popular parametric and non-parametric approaches are developed to assess the treatment effect for RCT. However, RCT may not be always feasible due to ethics, cost, and patient's preferences. With the availability of the observed data in natural healthcare setting, estimating treatment effect based on observational studies becomes more practical. In the observational studies, confounding covariates often exist, and the statistical methods developed for RCT may not be suitable for observational studies. In this project, I plan to extend some parametric and nonparametric methods to compare treatment effects among multiple groups. We used the superiority score as a measure of treatment effect between two groups. We will extend the parametric approaches such as ordinal logistic regression and nonparametric method such as adjusted U-statistic to compare the treatment effect in the presence of confounding covariates.

✉ soutikghosal@gmail.com

I5d. COMPARING MULTISTATE MODELING METHODS WITH APPLICATION TO ALZHEIMER'S DISEASE

Jacquelyn E. Neal*, *Vanderbilt University*

Dandan Liu, *Vanderbilt University*

When modeling progressive disease with multiple states, transition models with a Markov assumption are one of the most common methods used. When examining transitions between disease stages in Alzheimer's disease, however, the pathology of AD does not follow the Markov property. Other methods exist that do not rely on the Markov assumption, but their implementation is rare in AD literature. We have applied existing methodology from the literature to data from the National Alzheimer's Coordinating Center (NACC), specifically Markov multistate models, non-Markov multistate models, and partly conditional models. The ability and ease of these models to incorporate competing risk information, specifically death, and time-varying covariates will also be investigated. With progressive diseases linked to aging, the inclusion of competing risk of death is necessary to avoid bias, and because AD progression can span decades, the ability to include covariate information at different times is essential. We compare these differing methods when investigating transitions between disease stages and discuss the strengths and limitations of each method.

✉ jacquelyn.e.neal@vanderbilt.edu

I5e. ASYMPTOTIC CONFIDENCE INTERVAL CONSTRUCTION FOR PROPORTION RATIO BASED ON CORRELATED PAIRED DATA

Xuan Peng*, *State University of New York at Buffalo*

Changxing Ma, *State University of New York at Buffalo*

Song Liu, *Roswell Park Cancer Institute*

In ophthalmological and otolaryngology studies, measurements obtained from both organs (eyes or ears) of an individual are often highly correlated. Ignoring the intraclass correlation between paired measurement may yield biased

inferences. In this article, three different methods (maximum likelihood estimates based Wald-type confidence interval, profile likelihood confidence interval and asymptotic score confidence interval) are applied to this type of correlated bilateral data to construct confidence interval for proportion ratio, taking the intraclass correlation into consideration. The coverage probabilities and widths of the resulting methods are compared with existing methods in a Monte Carlo simulation study to evaluate their performance. A real data example from an ophthalmologic study is used to illustrate our methodology.

✉ xuanpeng@buffalo.edu

I6. POSTERS: SPATIAL AND TEMPORAL MODELING

I6a. ANALYZING SPATIAL LONGITUDINAL INCIDENCE PATTERNS USING DYNAMIC MULTIVARIATE POISSON MODELS

Yihan Sui*, *The Ohio State University*

Chi Song, *The Ohio State University*

Grzegorz Rempala, *The Ohio State University*

Statistical methods for multivariate, longitudinal, and continuous data have a long history and have been applied extensively in virtually all areas of modern science. In contrast, there is relatively little literature on count data models that account for both spatial and temporal dependence. We propose herein a hierarchical multivariate Poisson (MVP) model that simultaneously incorporates spatial correlations and temporal effects in a general broad setting. MVP allows for modeling the spatial/temporal dependent counts as a function of both location and time-varying covariates. To analyze temporal trends, we also add to MVP a broken-line regression model for detecting trend changes. Bayesian inference is adopted in a Markov chain Monte Carlo (MCMC) algorithm to estimate the parameters. In order to evaluate the fitness of the model, we propose a goodness-of-fit test for generalized linear model (GLM). The derivation of the test is based on standardized residuals. We show that the test statistic has an asymptotic normal distribution.

✉ sui.35@osu.edu

I6b. ONLINE SEQUENTIAL MONITORING OF DISEASE INCIDENCE RATES WITH AN APPLICATION TO THE FLORIDA INFLUENZA-LIKE ILLNESS DATA

Kai Yang*, *University of Florida*

Peihua Qiu, *University of Florida*

Online sequential monitoring of the incidence rates of chronic or infectious diseases is critically important for public health and stability of our society. Governments around the world have invested a great amount of resource in building global, national and regional disease reporting and surveillance systems. In these systems, conventional control charts, such as the cumulative sum and exponentially weighted moving average charts, are usually included for disease surveillance purpose. However, these charts require many assumptions on the observed data, including the ones of independent and identically normally distributed data when no disease outbreaks are present. These assumptions can hardly be valid in practice, making the results from the conventional control charts unreliable. Motivated by an application to monitor the Florida influenza-like illness data, we develop a new sequential monitoring approach in this paper, which can accommodate the dynamic nature of the disease incidence rates, spatio-temporal data correlation, and non-normality. It is shown that the new method is much more reliable to use in practice than the commonly used conventional charts.

✉ yklmy1994121@ufl.edu

I6c. POINTWISE TOLERANCE INTERVALS FOR AUTOREGRESSIVE MODELS, WITH AN APPLICATION TO HOSPITAL WAITING LISTS

Kedai Cheng*, *University of Kentucky*

Derek S. Young, *University of Kentucky*

Long waiting lists are a symbol of inefficiencies of hospital services. It is complex to understand how the lists grow due to the demand of a particular treatment relative to a hospitals capacity. Understanding the uncertainty of forecasting growth/decline of waiting lists could help hospital managers with capacity planning. We address this uncertainty through the use of statistical tolerance intervals, which are intervals that contain a specified proportion of the sampled population at a given confidence level. Tolerance intervals are available for numerous settings, including most classic parametric distributions, the nonparametric setting, and for some parametric models; however, there are no standard or well-accepted approaches for autoregressive models. This talk fills that gap and introduces an approach for establishing pointwise tolerance intervals for classic autoregressive models. Some theoretical developments of tolerance intervals in this setting will be discussed. A summary of simulation studies and an application to hospital waiting lists will be presented, which demonstrate the good coverage properties of this approach.

✉ kch268@g.uky.edu

I6d. CHARACTERIZING SPATIAL DEPENDENCE ON STREAM NETWORKS: BAYESIAN HIERARCHICAL MODEL APPROXIMATION

Yingying Liu*, *University of Iowa*

Stream network models are used to analyze the movement of substances or aquatic life along streams and rivers. In order to model stream network data based on stream distance and water flow, Ver Hoef and Peterson (2010) developed moving average models to construct two types of valid autocovariances: tail-up models and tail-down models. The goal is to distinguish between pure tail-up and pure tail-down covariance structures. For faster computation, we extend the idea of building conditional autoregressive (CAR) models as approximations to geostatistical models on a lattice has been introduced by Rue and Tjelmeland (2002) and use Bayesian hierarchical normal intrinsic conditional autoregressive (HNICAR) model to approximate stream network models.

✉ emilyingliu@gmail.com

I6e. MULTIVARIATE AIR POLLUTANT EXPOSURE PREDICTION IN SOUTH CAROLINA

Ray Boaz*, *Medical University of South Carolina*

Andrew Lawson, *Medical University of South Carolina*

John Pearce, *Medical University of South Carolina*

Air pollution is associated with adverse health outcomes ranging from increased respiratory incidence to increased mortality; however, the health impacts from exposure to multiple pollutants remain unclear. Large gaps in knowledge remain for developing models that address the decomposition of chemical mixtures in relation to health outcomes. Because air quality measurement predictions are greatly limited by missing data, this project focuses on the development of methods for improved estimation of pollutant concentrations when only sparse monitor networks are found. Specifically, a multivariate air pollutant statistical model to predict spatio-temporally resolved concentration fields for multiple pollutants simultaneously is developed and evaluated. The multivariate predictions allow monitored pollutants to inform the prediction of non-monitored pollutants. These methods utilize widely available data resources, meaning that the improved predictive accuracy of sparsely monitored pollutant concentrations can benefit future studies by improving estimation of health effects and saving resources needed for supplemental air pollutant monitoring campaigns.

✉ boaz@musc.edu

I6f. A BAYESIAN GLMM FOR MODELING SPATIALLY VARYING TRENDS IN DISEASE PREVALENCE WITH AN APPLICATION TO LYME DISEASE

Stella C. Watson*, *Clemson University*

Christopher S. McMahan, *Clemson University*

Andrew Brown, *Clemson University*

Robert Lund, *Clemson University*

This work considers the development of a Bayesian generalized linear mixed model, for the purposes of assessing disease trends throughout the conterminous United States (US). The model explicitly accounts for the spatial and temporal correlation structures that are omnipresent in such applications. Further, the model allows for spatial varying trends; i.e., it can account for the effect of increasing/decreasing disease prevalence within different geographic regions, while controlling for various confounding effects. The proposed approach was specifically designed to account for large spatio-temporal data sets, as the motivating example consists of 6 years of monthly data reported at the county level. In particular, in this work we consider modeling the prevalence of Lyme disease within the canine population across the US. The goal of this analysis was to identify regions of the country that are experiencing an increase or decrease in risk.

✉ stellaw@clemson.edu

I6g. MODELING HIGH DIMENSIONAL MULTICHANNEL BRAIN SIGNALS

Lechuan Hu*, *University of California, Irvine*

Norbert J. Fortin, *University of California, Irvine*

Hernando Ombao, *King Abdullah University of Science and Technology*

Our goal is to model and measure functional and effective (directional) connectivity in multichannel brain physiological signals (e.g., electroencephalograms, local field potentials). To model multichannel brain signals, our approach is to fit a vector autoregressive (VAR) model with potentially high lag order so that complex lead-lag temporal dynamics between the channels can be captured. Estimates of the VAR model will be obtained by our proposed hybrid LASSLE (LASSO+LSE) method which combines regularization (to control for sparsity) and least squares estimation (to improve bias and mean-squared error). Then we employ some measures of connectivity but put an emphasis on partial directed coherence (PDC) which can capture the directional connectivity between channels. The proposed modeling approach provided key insights into potential functional relationships among simultaneously recorded sites during performance of a complex memory task. It quantified patterns of connectivity and its evolution across trials.

✉ lechuanh@uci.edu

I6h. IDENTIFYING NON-STATIONARITY IN PM2.5 DATA VIA AN M-RA AND MIXTURE PRIORS

Marco Henry Benedetti*, *University of Michigan*

Veronica Berrocal, *University of Michigan*

Naveen Narisetty, *University of Illinois at Urbana-Champaign*

A challenge in air pollution epidemiological studies is the lack of information on ambient exposure for most subjects. To circumvent this problem and derive point-level estimates of air pollution, several methods have been proposed, including spatial statistical models that assume the spatial correlation decays at a constant rate throughout the domain. This assumption may not be appropriate for PM2.5, a mixture of pollutants that include both long-range contaminants and ones from more localized sources. To address this, building upon the M-RA model introduced by Katzfuss (JASA, 2016), we express the spatial field as a linear combination of multi-resolution basis functions, and provide the basis function weights with resolution-specific mixture priors. Simulation studies demonstrate that our model can detect regions with varying rates of spatial decay. Additionally, an application to daily average PM2.5

concentration indicates that: (i) the pattern of the spatial dependence of PM2.5 is non-homogeneous and (ii) out-of-sample predictions generated from our model are better than those obtained via ordinary kriging in terms of MSE and empirical coverage of prediction intervals.

✉ benedetm@umich.edu

I6i. TEMPORALLY DEPENDENT ACCELERATED FAILURE TIME MODEL FOR CAPTURING THE IMPACT OF EVENTS THAT ALTER SURVIVAL IN DISEASE MAPPING

Rachel M. Carroll*, *National Institute of Environmental Health Sciences, National Institutes of Health*

Shanshan Zhao, *National Institute of Environmental Health Sciences, National Institutes of Health*

Andrew B. Lawson, *Medical University of South Carolina*

The introduction of spatial and temporal frailties in survival models furnishes a way to represent unmeasured confounding in the outcome. Using a Bayesian accelerated failure time model, we can flexibly address a wide range of spatial and temporal options for structuring frailties as well as examine the benefits of using these different structures in certain settings. Our results suggest that it is important to include temporal frailties when a true temporal structure is present and including them when a true temporal structure is absent does not sacrifice model fit. Further, frailties can correctly recover the truth imposed on simulated data without affecting the fixed effect estimates. In the case study involving breast cancer mortality, the temporal frailty played an important role in representing the unmeasured confounding related to improvements in disease screenings as well as the impact of Hurricane Katrina. In conclusion, the incorporation of spatial and temporal frailties in survival analysis can lead to better fitting models and improved inference by representing spatially and temporally varying unmeasured risk factors and confounding that could impact survival.

✉ rachel.carroll@nih.gov

I7. HUMAN MICROBIOME ANALYSIS: NEW STUDY DESIGNS, NOVEL METHODS, AND PRACTICAL CONSIDERATIONS

METHODS FOR INFERRING GROWTH DYNAMICS OF GUT MICROBIOTA FROM METAGENOMICS SAMPLES

Hongzhe Li*, *University of Pennsylvania*

Metagenomic sequencing increased our understanding of the role of the microbiome in health and disease, yet it only provides a snapshot of a highly dynamic ecosystem. Based on the read coverage in metagenomes, one can potentially infer the growth rates of the species in the samples. However, since metagenomic samples includes many known and unknown microbial species, to fully describe the growth rates of all the species, we fist construct contigs and examine the read coverage distributions across samples. We present statistical methods for modeling such data across all samples. We illustrate the methods using metagenomic studies of IBD and normal gut microbiome samples and show a large difference in species growth dynamics in the disease samples. Differential growth dynamics analysis will also be discussed.

✉ hongzhe@pennmedicine.upenn.edu

IDENTIFYING HOST GENETIC VARIANTS ASSOCIATED WITH MICROBIOME COMPOSITION IN GENOME-WIDE ASSOCIATION STUDIES

Jianxin Shi*, *National Cancer Institute, National Institutes of Health*

Xing Hua, *National Cancer Institute, National Institutes of Health*

Identifying host genetic variants associated with human microbiome composition provides clues for characterizing microbiome variation and helps to elucidate biological mechanisms of genetic associations. Since a microbiota

functions as a community, it is best characterized by beta diversity. We develop a statistical framework and a computationally efficient software package, microbiomeGWAS, for identifying host genetic variants associated with microbiome beta diversity. We show that score statistics have positive skewness and kurtosis due to the dependent nature of the pairwise data, which makes P-value approximations based on asymptotic distributions unacceptably liberal. By correcting for skewness and kurtosis, we develop accurate P-value approximations with accuracy verified by extensive simulations. We exemplify our methods by analyzing a set of 147 genotyped subjects with 16S rRNA microbiome profiles from non-malignant lung tissues. Correcting for skewness and kurtosis eliminated the dramatic deviation in the quantile-quantile plots. We provided evidence that six established lung cancer risk SNPs were collectively associated with microbiome composition.

✉ jianxin.shi@nih.gov

PRACTICAL ISSUES IN ANALYZING LONGITUDINAL HUMAN MICROBIOME DATA

Snehalata Huzurbazar*, *West Virginia University*

Eugenie Jackson, *West Virginia University*

As longitudinal human microbiome data becomes available, we approach the literature and the software from a practioner's point of view. First, we survey the literature to assess whether the methods being used for longitudinal analyses are useful and whether they help answer questions motivating the studies. Next, we survey and use software that is available for longitudinal analyses to assess how well the software works given the characteristics of human microbiome data. Some examples of recent software include BioMiCo and CORAL. We use two data sets with different characteristics with respect of sparseness and n relative to p to compare the software.

✉ snehalata.huzurbazar@hsc.wvu.edu

KERNEL MACHINE REGRESSION METHODS FOR CORRELATED MICROBIOME COMMUNITY DATA

Ni Zhao*, *Johns Hopkins University*

Haotian Zheng, *Tsinghua University*

Xiang Zhan, *The Pennsylvania State University*

In the past few years, research interest in the human microbiome increases greatly. Many studies now collect data that are correlated, such as repeated measurements of microbiome from the same subjects in longitudinal studies, measurements that are clustered based on location, shared resources, or family membership. However, existing methods mainly focus on cross-sectional studies with independent subjects, and few methods are developed to analyze correlated microbiome data. Here, we propose a kernel machine regression approach to evaluate the association between the microbiome composition and a phenotype of interest, considering the correlations within the microbiome and the phenotype. Kernels, which measure the pairwise similarities in microbiome profiles, are constructed by converting the well-accepted distance measurements, allowing for incorporation of the phylogenetic information. Generalized linear mixed model, which takes into account the intrasubject correlations, is used for parameter estimation; and a variance component score test is developed for association testing. We evaluate the performance of our approach via extensive simulations and a real study.

✉ nzhao10@jhu.edu

I8. RECENT INNOVATION IN NETWORK META-ANALYSIS

BAYESIAN INFERENCE FOR NETWORK META-REGRESSION USING MULTIVARIATE RANDOM EFFECTS WITH APPLICATIONS TO CHOLESTEROL LOWERING DRUGS

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

Ming-Hui Chen, *University of Connecticut*

Arvind Shah, *Merck & Co., Inc.*

Hao Li, *University of Connecticut*

Sunduk Kim, *Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health*

Jianxin Lin, *Merck & Co., Inc.*

Andrew Tershakovec, *Merck & Co., Inc*

Low-density lipoprotein cholesterol (LDL-C) has been identified as a causative factor for atherosclerosis and related coronary heart disease, and as the main target for cholesterol-lowering and lipid-lowering therapy. Statin drugs inhibit cholesterol synthesis in the liver and are typically the first line of therapy to lower elevated levels of LDL-C. On the other hand, a different drug, Ezetimibe, inhibits the absorption of cholesterol by the small intestine and provides a different mechanism of action. To synthesize the results from different clinical trials, we examine treatment level (aggregate) network meta-data from 29 double-blind, randomized, active or placebo-controlled statins +/- Ezetimibe clinical trials on adult treatment-naïve patients with primary hypercholesterolemia. We propose a new approach to carry out Bayesian inference for arm-based network meta-regression. The proposed approach is especially useful when some treatment arms are involved in only a single trial. The proposed methodology is further applied to analyze the network meta-data from 29 trials with 11 treatment arms.

✉ ibrahim@bios.unc.edu

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

Haitao Chu*, *University of Minnesota*

Xiaoye Ma, *University of Minnesota*

Qinshu Lian, *University of Minnesota*

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

Yong Chen, University of Pennsylvania

To compare the accuracy of multiple tests in a single study, three designs are commonly used including: 1) the multiple test comparison design; 2) the randomized design and 3) the non-comparative design. The increasing number of available diagnostic instruments for a disease condition has generated the need to develop efficient meta-analysis methods to simultaneously compare multiple instruments. However, existing meta-analysis of diagnostic tests (MA-DT) have been focused on evaluating a single test by comparing it with a reference test. In this paper, we propose a Bayesian hierarchical model for network meta-analysis of diagnostic tests (NMA-DT) to simultaneously compare multiple diagnostic tests. We develop a missing data framework for NMA-DT and offer four important promises over MA-DT: 1) it combines information from studies with all three designs; 2) it pools both studies with or without a gold standard; 2) it allows different sets of candidate tests in different studies; and 4) it accounts for potential heterogeneity across studies and complex correlation structure among multiple diagnostic tests. We illustrate our method through two case and simulation studies.

✉ chux0051@umn.edu

N-OF-1 TRIALS FOR MAKING PERSONALIZED TREATMENT DECISIONS

Christopher H. Schmid*, *Brown University*

N-of-1 trials, single-participant multiple-crossover studies to determine the comparative effectiveness of two or more treatments, can enable patients to create personalized protocols to guide medical care. An individual selects treatments and outcomes of interest, carries out the trial, and then makes a final treatment decision with or without a clinician based on results of the trial. Established in a clinical environment, an N-of-1 practice provides data on multiple trials from different patients. Such data can be combined using meta-analytic techniques to inform both individual and population treatment effects. When patients undertake trials with different treatments, the data form a treatment network and suggest use of network meta-analysis methods. This talk will discuss clinical research projects using N-of-1 trials and will discuss design and analytic challenges deriving from use of the N-of-1 design for personalized decision-making. These include defining treatments, presenting results, assessing model assumptions and combining information from different patients to improve estimates of individual effects.

✉ christopher_schmid@brown.edu

NETWORK META-REGRESSION FOR ORDINAL OUTCOMES: APPLICATIONS IN COMPARING CROHN'S DISEASE TREATMENTS

Ming-Hui Chen*, *University of Connecticut*

Yeongjin Gwon, *University of Connecticut*

May Mo, *Amgen Inc.*

Juan Li, *Eli Lilly and Company*

H. Amy Xia, *Amgen Inc.*

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

In the U.S., there are approximately 780,000 Crohn's disease patients and 33,000 new cases are added each year. In this paper, we propose a new network meta-regression approach for modeling ordinal outcomes in order to assess the efficacy of treatments for Crohn's disease. Specifically, we develop regression models based on aggregate trial-level covariates for the underlying cut-off points of the ordinal outcomes as well as for the variances of the random effects to capture heterogeneity across trials. Our proposed models are particularly useful for indirect comparisons of multiple treatments that have not been compared head-to-head within the network meta-analysis framework. Moreover, we introduce Pearson residuals to detect outlying trials and construct an invariant test statistic to evaluate goodness-of-fit in the setting of ordinal outcome meta-data. A detailed case study demonstrating the usefulness of the proposed methodology is carried out using aggregate ordinal outcome data from 16 clinical trials for treating Crohn's disease.

✉ ming-hui.chen@uconn.edu

I9. STATISTICAL ANALYSIS OF TRACKING DATA FROM PERSONAL WEARABLE DEVICES

EMERGING BIOSTATISTICAL PROBLEMS IN WEARABLE AND IMPLANTABLE TECHNOLOGY (WIT)

Ciprian M. Crainiceanu*, *Johns Hopkins University*

The talk will focus on describing emerging data structures obtained from continuous monitoring of human health using wearable devices, such as accelerometers and heart monitors, and implantable devices, such as glucometers. In particular, I will describe the multi-scale nature of the data and introduce methodological problems associated with micro- and macro-scale WIT data and its association with health outcomes.

✉ ccrainic@jhsph.edu

USE OF ACCELEROMETERS IN CLINICAL TRIALS

John W. Staudenmayer*, *University of Massachusetts, Amherst*

This talk will describe a recent pharmaceutical trial that investigated the effectiveness of a drug to increase physical activity in a specific clinical population. The multi-site trial used an accelerometer to estimate the change in physical activity via a placebo controlled cross over design. The talk will describe the trial's measurement protocols and present the results. Additionally, we will review statistical methods to estimate aspects of physical activity and inactivity using an accelerometer. This will include simpler linear regression methods and more sophisticated statistical learning methods. We will also describe how we assessed the clinical relevance of the magnitude of the changes that were observed in the accelerometer outcome.

✉ jstauden@math.umass.edu

STATISTICAL ANALYSIS OF SOCIAL AND BEHAVIORAL MARKERS FROM SMARTPHONE DATA

Jukka-Pekka Onnela*, *Harvard University*

Recent advances in biomedicine and technology are beginning to change the priority in biomedical research towards phenotyping. We believe that the ubiquity and capability of smartphones to collect social and behavioral data can contribute to the so-called phenotyping challenge via objective measurement, especially in neuropsychiatric conditions, where phenotyping and measurement of patient-centered outcomes outside the clinic remains very challenging. We have defined digital phenotyping as the "moment-by-moment quantification of the individual-level human phenotype in situ using data from personal digital devices," in particular smartphones. In a typical study, we collect approximately 1GB of high-dimensional, longitudinal data per patient-month, and while data collection at scale is getting easier, data analysis is increasingly identified as a major bottleneck in this area of research. I will talk about some of our recent work on the statistical analysis of social and behavioral markers from smartphone data, focusing on the construction of meaningful summary statistics, dealing with missing data, and longitudinal data analysis.

✉ onnela@hsph.harvard.edu

QUANTIFYING HERITABILITY OF PHYSICAL ACTIVITY PATTERNS BASED USING FUNCTIONAL ACE MODELS

Haochang Shou*, *University of Pennsylvania*

Joanne Carpenter, *University of Sydney*

Kathleen Merikangas, *National Institute of Mental Health, National Institutes of Health*

Ian Hickie, *University of Sydney*

Twin studies provide unique opportunities to understand heritability of certain quantitative traits. The conventional ACE models via structural equation modeling have been used to segment and quantify the genetic and environmental influences in the sample. However, the emergence of complex measures generated by novel technologies such as wearable sensor devices has posed challenges on directly applying such techniques to time-dependent measures. We extended the traditional ACE model for a single univariate trait to functional outcomes. In particular, the method simultaneously: 1) handle various levels of dependency in the data; 2) identify interpretable traits via dimensionality reduction based on principal components; and 3) estimate relative variances that are attributed by additive genetic, shared environmental and unique environmental effects. Within-family similarities of those complex measures could also be effectively quantified. The methods have been applied to define heritable features in physical activity on the actigraphy subsample of the Brisbane adolescent twin study.

✉ hshou@pennmedicine.upenn.edu

20. TEACHING DATA SCIENCE AT ALL LEVELS

DATA SCIENCE AS A GATEWAY TO STATISTICS

Mine Cetinkaya-Rundel*, *Duke University and RStudio, Inc.*

In this talk we will discuss a data science course designed to serve as a gateway to the discipline of statistics, the statistics major, and broadly to quantitative studies. The course is intended for an audience of Duke University students with little to no computing or statistical background, and focuses on data wrangling, exploratory data analysis, data visualization, and effective communication. Unlike most traditional introductory statistics courses, this course approaches statistics from a model-based perspective and introduces simulation-based and Bayesian inference later in the course. A heavy emphasis is placed on reproducibility (with R Markdown) and version control and collaboration (with git/GitHub). In this talk we will discuss in detail the course structure, logistics, and pedagogical considerations as well as give examples from the case studies used in the course. We will also share student feedback, assessment of the success of the course in recruiting students to the statistical science major, and our experience of growing the course from a small seminar course for first-year undergraduates to a larger course open to the entire undergraduate student body.

✉ mine@stat.duke.edu

TEACHING DATA SCIENCE FOR LIFE SCIENCES

Michael I. Love*, *University of North Carolina, Chapel Hill*

Biologists and biomedical researchers find their fields have rapidly advanced toward a state where experiments produce large data outputs they are not formally trained to evaluate. These large datasets must be processed, normalized and appropriately modeled before making scientific inferences. I will discuss various forums through which biologists and biomedical researches are updating their data science skills for working with and publishing on these data, and what informs their choice between collaborating with quantitative researchers and developing data science skills within the "wet lab". I will also discuss considerations regarding statistical content of short courses and MOOCs available to these scientists to augment their data analytic skills for modern datasets.

✉ michaelisaiahlove@gmail.com

MAKE INTERACTIVE WEB TUTORIALS WITH learnr AND R

Garrett Grolemund*, *RStudio, Inc.*

The learnr R package provides a new multimedia approach for teaching statistics and programming with R. With learnr, teachers can combine text, diagrams, videos, pacing cues, code exercises, multiple choice questions, automated grading software and more to create an interactive, self-paced tutorial. Learnr is based on the familiar R Markdown format, which makes it easy to write learnr tutorials and to host them online. This talk will demonstrate the learnr package and examine several best practices for teaching in a multi-media, self paced format format, a format that may be new to many teachers.

✉ garrett@rstudio.com

TEACHING SURVEY AND DATA SCIENCE OUTSIDE REGULAR CLASSROOM SETTINGS

Frauke Kreuter*, *University of Maryland and University of Mannheim*

Over the last three years we experimented with various events and teaching activities to bring non-STEM students and practitioners up to speed on Survey and Data Science. This talk will highlight three approaches. First, the DataFest, a Data Analysis challenge designed for students to learn and apply data analysis skills during a weekend to create insights out of novel data. Second, the Advanced Data Analytics training provides training in Data Science

for federal, state and local government program agency employees (coleridgeinitiative.org). Third, an International Professional Training Program in Survey and Data Science (survey-data-science.net). In all instances the the majority of the participants are neither computer scientists nor do they have any extensive training in statistics. We found the task-oriented approach with strong peer-to-peer elements to show remarkable successes in brining non-technical people of all ages into a situation where they can critically analyze complex data, and learn how to self-enhance their skillset.

✉ fkreuter@umd.edu

21. RICH DATA VISUALIZATIONS FOR INFORMATIVE HEALTH CARE DECISIONS

FLEXIBLE AND INTERPRETABLE REGRESSION IN HIGH DIMENSIONS

Ashley Petersen*, *University of Minnesota*

Daniela Witten, *University of Washington*

In recent years, it has become quick and inexpensive to collect and store large amounts of data in a number of fields. With big data, the traditional plots used in exploratory data analysis can be limiting, given the large number of possible predictors. Thus, it can be helpful to fit sparse regression models, in which variable selection is adaptively performed, to explore the relationships between a large set of predictors and an outcome. For maximal utility, the functional forms of the covariate fits should be flexible enough to adequately reflect the unknown relationships and interpretable enough to be useful as a visualization technique. We will provide an overview of recent work in the area of sparse additive modeling that can be used for visualization of relationships in big data. In addition, we present recent novel work that fuses together the aims of these previous proposals in order to not only adaptively perform variable selection and flexibly fit included covariates, but also adaptively control the complexity of the covariate fits for increased interpretability.

✉ pete6459@umn.edu

VISUALISING MODEL STABILITY INFORMATION FOR BETTER PROGNOSIS BASED NETWORK-TYPE FEATURE EXTRACTION

Samuel Mueller*, *University of Sydney*

Connor Smith, *University of Sydney*

Boris Guennewig, *University of Sydney*

In this talk, we present our latest findings to deliver new statistical approaches to identify various types of interpretable feature representations that are prognostically informative in classifying complex diseases. Identifying key features and their regulatory relationships which underlie biological processes is the fundamental objective of much biological research; this includes the study of human disease, with direct and important implications in the development of target therapeutics. We present new and robust ways to visualise valuable information from the thousands of resamples in modern selection methods that use repeated subsampling to identify what features predict best disease progression. We show that using subtractive lack-of-fit measures scales up well to large dimensional situations, making aspects of exhaustive procedures available without its computational cost.

✉ samuel.mueller@sydney.edu.au

VISUALIZATIONS FOR JOINT MODELING OF SURVIVAL AND MULTIVARIATE LONGITUDINAL DATA IN HUNTINGTON'S DISEASE

Jeffrey D. Long*, *University of Iowa*

We discuss the joint modeling of survival data and multivariate longitudinal data in several Huntington's disease (HD) data sets. HD is an inherited disorder caused by a cytosine-adenine-guanine (CAG) expansion mutation, and it is characterized primarily by motor disturbances, such as chorea. Development of new methods of genetic analysis allow researchers to find genetic variants other than CAG that modify the timing of motor diagnosis. The goal of the analysis was to compute an individual-specific residual phenotype that might be used in subsequent genetic analysis. A martingale-like residual is defined that represents the deviance of a participant's observed status at the time of motor diagnosis or censoring and their concurrent model-predicted status. It is shown how the residual can be used to index the extent to which an individual is early or late (or on time) for motor diagnosis. A Bayesian approach to parameter estimation is taken, and methods of external validation are illustrated based on the time-dependent area under the curve (AUC). Visualization of residuals for scientific importance and statistical characteristics are shown.

✉ jeffrey-long@uiowa.edu

DISCORDANCY PARTITIONING FOR VALIDATING POTENTIALLY INCONSISTENT PHARMACOGENOMIC STUDIES

J. Sunil Rao*, *University of Miami*

Hongmei Liu, *University of Miami*

The Genomics of Drug Sensitivity (GDSC) and Cancer Cell Line Encyclopedia (CCLE) are two major studies that can be used to mine for therapeutic biomarkers for cancers. Model validation using the two datasets however has proved elusive and has put into some question the usefulness of such large scale pharmacogenomic assays. While the genomic profiling seems consistent, the drug response data is not. We present a partitioning strategy based on a data sharing concept which directly acknowledges a potential lack of concordance between datasets and in doing so, also allows for extraction of new and reproducible signal. We show both significantly improved test set prediction accuracy over existing methods and develop some new visualization tools for signature validation.

✉ jrao@biostat.med.miami.edu

22. MODERN RANDOMIZED TRIAL DESIGNS

THE IMP: INTERFERENCE MANIPULATING PERMUTATION

Michael Baiocchi*, *Stanford University*

This talk provides a framework for randomization in situations where the intervention level for one unit of observation has the potential to impact other units' outcomes. The goal of the interference manipulating permutation (IMP) is to reduce interference between units, improving the data quality in anticipation of using one of several forms of inference developed to obtain traditional causal estimates in the presence of interference. This approach may be particularly of interest to investigators interested in improving decision-making in the prevention of infectious disease or deploying behavioral interventions. The framework is motivated by two cluster-randomized trials (CRTs) of a behavioral health intervention delivered in schools situated within the informal settlements of Nairobi, Kenya. Interviews collected from the pilot study indicated that the young girls felt motivated to share the skills gained from the intervention with their friends and family. IMP was developed and deployed for the formal CRT study of the intervention. This proposed framework draws upon earlier work by Moulton (2004) and Tukey (1993).

✉ mike.baiocchi@gmail.com

TRANSLATING CLINICAL TRIAL RESULTS TO A TARGET EHR POPULATION USING MACHINE LEARNING AND CAUSAL INFERENCE

Benjamin A. Goldstein*, *Duke University*

Matt Phelan, *Duke Clinical Research Institute*

Neha Pagidipati, *Duke University*

While randomized clinical trials (RCT) are the gold standard for estimating treatment effects, their results can be misleading if there is treatment heterogeneity. The effect estimated in the trial population may differ from the effect in some different population. In this presentation we combine methodology from machine learning and causal inference to translate the results from a RCT to a target Electronic Health Record (EHR) based population. Using RCT data we build a random forests prediction model among those that received two different treatments. We then use the principals of Causal Random Forests to estimate a predicted disease outcome under both treatment conditions within the target population. We estimate each individual's treatment effect and average over the target sample to obtaine the population avergae treatment effect. Using real data we show that we obtain internally consist estimates within the original trial, and new inference within the target sample.

✉ ben.goldstein@duke.edu

EVALUATING EFFECTIVENESS AND SAFETY OF LOW AND HIGH DOSE ASPIRIN: A PRAGMATIC TRIAL APPROACH

Zhen Huang*, *Duke Clinical Research Institute*

Jennifer White, *Duke Clinical Research Institute*

Frank Rockhold, *Duke Clinical Research Institute*

Aspirin is a mainstay therapy for patients with atherosclerotic cardiovascular disease. Although millions of Americans take aspirin every day or every other day for secondary prevention, the optimal dose has not been established. ADAPTABLE is a pragmatic trial attempt to answer this question. In this study, participants are identified through electronic health record (EHR) computable phenotype. Sign of consent, randomization, and follow up are carried out by participants in online patient portal. Outcomes information is collected through EHR data in PCORnet DataMarts, complemented by insurance data and National Death Index. In this talk, we will highlight the unique features of the study design and discuss potential challenges, including the availability and reliability of subject self-reported and EHR data, the concordance among multiple data sources, validation of endpoints in lieu of clinical adjudication committee, and the role of the independent data monitoring committee during the study.

✉ zhen.huang@duke.edu

CAUSAL ANALYSIS OF SELF-TRACKED TIME SERIES DATA USING A COUNTERFACTUAL FRAMEWORK FOR N-OF-1 TRIALS

Eric J. Daza*, *Stanford Prevention Research Center*

Many types of personal health data form a time series (e.g., wearable-device data, regularly monitored clinical events, chronic conditions). Causal analyses of such n-of-1 (i.e., single-subject) observational studies (N1OSs) can be used to discover possible cause-effect relationships to then self-test in an n-of-1 randomized trial (N1RT). This talk introduces and characterizes the average period treatment effect (APTE) as the N1RT estimand of interest, and builds a basic analytical framework that can accommodate autocorrelation and time trends in the outcome, effect carryover from previous treatment periods, and slow onset or decay of the effect. The APTE is loosely defined as a contrast of averages of potential outcomes the individual can theoretically experience under different treatment levels during a given treatment period. Two common causal inference methods are specified within the N1OS context, and used to search for estimable and interpretable APTEs using six years of the author's self-tracked weight and exercise data. Both the preliminary findings and the challenges faced in conducting N1OS causal discovery are reported.

✉ ericjdaza@stanford.edu

23. CLINICAL TRIAL METHODS

BAYESIAN CONTINUOUS MONITORING FOR PHASE I COHORT EXPANSION AND PHASE II CANCER CLINICAL TRIALS

Youjiao Yu*, *Baylor University*

Bayesian methods are widely used in cancer clinical trials for its flexibility in continuously monitoring a trial to make timely decisions, as well as the capability to incorporate a priori information in the model to make more informed decisions. We propose a Bayesian design based on two posterior probabilities to monitor both efficacy and futility for phase I cohort expansion and phase II cancer clinical trials. Dynamic stopping boundaries are proposed to increase the power of the design. Advantages of our design include flexibility in continuously monitoring a clinical trial, straightforward interpretation of efficacy and futility for interim data, and high statistical power or lower expected sample size compared to other similar cancer clinical trial designs.

✉ Youjiao_Yu@baylor.edu

USING MULTI-STATE MODELS IN CANCER CLINICAL TRIALS

Jennifer G. Le-Rademacher*, *Mayo Clinic*

Ryan A. Peterson, *University of Iowa*

Terry M. Therneau, *Mayo Clinic*

Sumithra J. Mandrekar, *Mayo Clinic*

Time-to-event endpoints are common in cancer trials and are commonly analyzed with Kaplan-Meier curves, logrank tests and Cox models. However, in trials with complex disease process and/or treatment options, multistate models (MSM) add important insights.This talk will focus on simple Aalen-Johansen estimates - the multistate analog of the Kaplan-Meier - via the analysis of a leukemia trial. The canonical path for a subject in the trial is a conditioning regimen (A or B), which leads to a complete response (CR), followed by consolidation therapy, and eventually followed by relapse and death. While standard survival methods look only at A vs. B in terms of overall survival, MSM can track all the intermediate states in a manner that is simple to compute and interpret. In our leukemia trial, MSM provides significant insights that the survival advantage observed in the experimental treatment results from its ability to both induce a faster CR and prolong survival once a patient achieved CR. Our goal is to encourage the use of MSM in cancer trials as they complement standard survival methods and may facilitate a better understanding of cancer disease process and its treatments.

✉ Le-Rademacher.Jennifer@mayo.edu

CLARIFYING COMMON MISCONCEPTIONS ABOUT COVARIATE ADJUSTMENT IN RANDOMIZED TRIALS

Bingkai Wang*, *Johns Hopkins Bloomberg School of Public Health*

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

There is much variation in how baseline variables are used in the primary analysis of randomized trials. Some of this variation is due to misunderstandings about the benefits, limitations, and interpretation of statistical methods that adjust for prognostic baseline variables, called covariate adjustment. We aim to clarify some of these misunderstandings through analytic arguments, simulation studies, and clinical applications using data from completed randomized trials of drugs for mild cognitive impairment, schizophrenia, and depression, respectively. We untangle some counter-intuitive properties of covariate adjustment, e.g., that it simultaneously reduces conditional bias and unconditional variance; it has greater added value in large trials; it can increase power even when there is perfect balance across arms in the baseline variables; it can reduce sample size even when there is no treatment effect. We provide visualizations of how the conditional bias reduction due

to covariate adjustment leads directly to a gain in unconditional precision. We also show how missing data, treatment effect heterogeneity and model misspecification impact the gains from such adjustment.

✉ bwang51@jhu.edu

MMRM ESTIMATES CONSIDERATION FOR LONGITUDINAL DATA IN CLINICAL TRIALS

Zheng (Jason) Yuan*, *Vertex Pharmaceuticals*

Chenkun Wang, *Vertex Pharmaceuticals*

Bingming Yi, *Vertex Pharmaceuticals*

When analyzing repeated measurement (longitudinal) data in clinical trials, it is common to implement Mixed-effect Model Repeat Measurement (MMRM) model by SAS PROC MIXED to estimate the LS means. However, caution needs to be taken when categorical covariates are included in MMRM models as the LS means obtained from the models could be deviated from what you want in randomized clinical trials. One common issue is the LS means estimates sometimes give very different numbers from the naïve raw means if there are categorical covariates in the MMRM model. Another issue is that the MMRM model gives different estimates of both within-treatment and between-treatment effects when adding the interaction term between covariates and treatment group, as compare to models without this interaction term. We explore and evaluate these issues by both simulations and real data examples to find out the root cause of these issues and then propose the recommended approach of estimating the LS means by MMRM model for various real world scenarios.

✉ jason_yuan@vrtx.com

SURROGATE ENDPOINT EVALUATION: META-ANALYSIS, INFORMATION THEORY, AND CAUSAL INFERENCE

Geert Molenberghs*, *I-BioStat, Hasselt University and Katholieke Universiteit Leuven*

Surrogate endpoints have been studied by Prentice (1989), who presented a definition of validity as well as a formal set of criteria that are equivalent if both the surrogate and true endpoints are binary. Freedman, Graubard, and Schatzkin (1992) supplemented these criteria with the proportion explained which, conceptually, is the fraction of the treatment effect mediated by the surrogate. Noting operational difficulties with the proportion explained, Buyse and Molenberghs (1998) proposed instead to use jointly the within-treatment partial association of true and surrogate responses, and the treatment effect on the surrogate relative to that on the true outcome. In a multi-center setting, these quantities can be generalized to individual-level and trial-level measures of surrogacy. Buyse et al. (2000) therefore have therefore proposed a meta-analytic framework to study surrogacy at both the trial and individual-patient levels. Various others paradigms exist. More recently, information theory and causal-inference methods have usefully been applied. Alonso et al. (2017) gives a unified overview. We present an overview of these developments.

✉ geert.molenberghs@uhasselt.be

MILESTONE PREDICTION FOR TIME-TO-EVENT ENDPOINT MONITORING IN CLINICAL TRIALS

Fang-Shu Ou*, *Mayo Clinic*

Martin A. Heller, *Alpha Statistical Consulting*

Qian Shi, *Mayo Clinic*

Predicting the times of milestone events, i.e. interim and final analysis in clinical trials, helps resource planning. We investigate several easily implemented methods, in both frequentist and Bayesian frameworks, for predicting when a milestone event is achieved. We show that it is beneficial

to combine multiple prediction models to craft a better predictor via prediction synthesis. Furthermore, a Bayesian approach provides a better measure of the uncertainty involved in the prediction of milestone events. We compare the methods through two simulations; one where the model has been correctly specified and one where the models are a mixture of 3 incorrectly specified model classes. We then apply the method on a real clinical trial data, NCCTG N0147. The performance using Bayesian prediction synthesis is very satisfactory, i.e. the predictions are within 20 days of the actual milestone (interim at 50% of event) time after 20% of events were observed. In summary, the Bayesian prediction synthesis methods automatically perform well even when the model is incorrectly specified or data collection is far from homogeneous.

✉ ou.fang-shu@mayo.edu

24. ENVIRONMENTAL AND ECOLOGICAL APPLICATIONS

MODELING HOURLY SOIL TEMPERATURE MEASUREMENTS

Nels G. Johnson*, *U.S. Forest Service, Pacific Southwest Research Station*

David R. Weise, *U.S. Forest Service, Pacific Southwest Research Station*

Stephen S. Sackett, *U.S. Forest Service, Pacific Southwest Research Station*

Sally M. Haase, *U.S. Forest Service, Pacific Southwest Research Station*

Microbiological activity depends on the temperature of soil. Solar energy striking the earth's surface is either reflected or absorbed depending on the characteristics of the surface. We propose a Fourier-basis function approach for handling day/night and seasonal effects on hourly soil temperature. This approach uses interaction effects of basis functions to model increased variation (i.e., amplitude) in day/night effects over season. We illustrate the model on a hourly soil temperatures collected from a split-plot experiment in Chimney Spring, AZ which investigates the effect of burn regime, over story type, and soil depth on soil temperature.

✉ nelsjohnson@fs.fed.us

EFFICIENT ESTIMATION FOR NON-STATIONARY SPATIAL COVARIANCE FUNCTIONS WITH APPLICATION TO CLIMATE MODEL DOWNSCALING

Yuxiao Li*•, *King Abdullah University of Science and Technology*

Ying Sun, *King Abdullah University of Science and Technology*

Spatial processes exhibit non-stationarity in many climate and environmental applications. Convolution-based approaches are often used to construct non-stationary covariance functions in the Gaussian random field. Although convolution-based models are highly flexible, they are not easy to fit even when datasets are moderate in size, and their computation becomes extremely expensive when large datasets are large. Most existing efficient methods rely on fitting an anisotropic but stationary model locally and reconstructing the spatially varying parameters. In this paper, we propose a new estimation procedure to approximate a class of non-stationary Mate?rn covariances by the local-polynomial fitting of the covariance parameters. The proposed method allows for efficient estimation of a richer class of non-stationary covariance functions with the local-stationary model as a special case. We also implement algorithms for fast high-resolution simulation of non-stationary Gaussian random fields with application to climate model downscaling of precipitations.

✉ yuxiao.li@kaust.edu.sa

MODELING EXPOSURES TO POLLUTANTS AND INFERTILITY IN COUPLES: A KERNEL MACHINE REGRESSION APPROACH

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

In epidemiological studies of environmental pollutants in relation to human infertility, it is common that concentrations of many exposures are collected in both male and female partners. Such a couple-based study poses some challenges in analysis, especially when the total effect of chemical mixtures is of interest. The kernel machine regression can be applied to model such effects, while accounting for the highly-correlated structure within and across the exposures. However, it does not consider the partner-specific structure in these study data. We develop a weighted kernel machine regression method to model the joint effect of partner-specific exposures, in which a linear weight procedure is used to combine both partners concentrations. The proposed method reduces the number of exposures and provides an overall importance index of partners exposures in infertility risk. Simulation studies demonstrate the good performance of the method and application of the proposed method to a prospective infertility study suggests that male partner's exposure to polychlorinated biphenyls contributes more toward infertility.

✉ chenzhe@mail.nih.gov

CAUSAL KERNEL MACHINE MEDIATION ANALYSIS FOR ESTIMATING DIRECT AND INDIRECT EFFECTS OF AN ENVIRONMENTAL MIXTURE

Katrina L. Devick*, *Harvard School of Public Health*

Jennifer F. Bobb, *Group Health Research Institute*

Maitreyi Mazumdar, *Boston Children's Hospital*

Birgit Claus Henn, *Boston University School of Public Health*

David C. Bellinger, *Boston Children's Hospital*

David C. Christiani, *Harvard School of Public Health*

Robert O. Wright, *Icahn School of Medicine at Mount Sinai*

Brent A. Coull, *Harvard School of Public Health*

Linda Valeri, *McLean Hospital*

New statistical methodology is needed to formalize the natural direct effect (NDE), natural indirect effect (NIE), and controlled direct effect (CDE) of a mixture of exposures on an outcome through an intermediate variable. We implemented Bayesian Kernel Machine Regression (BKMR) models to obtain posterior samples of the NDE, NIE and CDE, through simulation of counterfactuals. This method allows for nonlinear effects and interactions between the co-exposures, mediator and covariates. We applied this methodology to quantify the contribution of birth length as a mediator between in utero co-exposure of arsenic, manganese and lead, and children's neurodevelopment, in a prospective birth cohort in rural Bangladesh. Upon hypothetical intervention to fix birth length at the 75th percentile value of 48cm, the direct effect was not significant, suggesting, targeted interventions on fetal growth can block part of the adverse effect of metals on neurodevelopment (CDE: -0.07, 95% CI: -0.30, 0.17). Our extension of causal mediation methodology that allows for a mixture of exposures is important for environmental health applications.

✉ khartzler@fas.harvard.edu

USING DEEP Q-LEARNING TO MANAGE FOOT AND MOUTH DISEASE OUTBREAKS

Sandya Lakkur*, *Vanderbilt University*

Christopher Fonnesbeck, *Vanderbilt University*

Deep Q-learning has advanced the field of artificial intelligence and machine learning. Perhaps one of the most notable achievements with this method was teaching an agent to play Atari, along with various other video games, and play better than a human. Deep Q-learning is only

beginning to breach the field of biostatistics. This analysis uses deep Q-learning to manage a foot-and-mouth disease outbreak. The management is specifically concerned with answering the question, "which farm should be culled next?" This question prompts the agent to choose between N actions, where each action is a choice of which of the N farms to cull. This approach has shown promise in managing outbreaks on a small scale, and will eventually help policy makers construct general rules for types of farms to cull or pre-emptively cull to manage the disease outbreak.

✉ sandya.s.lakkur@vanderbilt.edu

IDENTIFYING EPIGENETIC REGIONS EXHIBITING CRITICAL WINDOWS OF SUSCEPTIBILITY TO AIR POLLUTION

Michele S. Zemplenyi*, *Harvard University*

Mark J. Meyer, *Georgetown University*

Brent A. Coull, *Harvard University*

Growing evidence supports an association between prenatal exposure to air pollution and adverse child health outcomes, including asthma and cardiovascular disease. Depending on the time and dose of exposure, epigenetic markers may be altered in ways that disrupt normal tissue development. Bayesian distributed lag models (BDLMs) have previously been used to characterize the time-varying association between methylation level at a given probe and air pollution exposure over time. However, by modeling probes independently, BDLMs fail to incorporate correlations between nearby probes. Instead, we use a function-on-function regression model to identify time periods during which there is an increased association between air pollution exposure and methylation level at birth. By accommodating both temporal correlations across pollution exposures and spatial correlations across the genome, this framework has greater power to detect critical windows of susceptibility to an exposure than do methods that model probes or exposure data independently. We compare the BDLM and function-on-function models via simulation, as well as with data from the Project Viva birth cohort.

✉ mzemplenyi@gmail.com

COMBINING SATELLITE IMAGERY AND NUMERICAL MODEL SIMULATION TO ESTIMATE AMBIENT AIR POLLUTION: AN ENSEMBLE AVERAGING APPROACH

Nancy Murray*, *Emory University*

Howard H. Chang, *Emory University*

Yang Liu, *Emory University*

Heather Holmes, *University of Nevada, Reno*

Ambient fine particulate matter less than 2.5 μm in aerodynamic diameter (PM2.5) has been linked to various adverse health outcomes and has, therefore, gained interest in public health. However, the sparsity of air quality monitors greatly restricts the spatio-temporal coverage of PM2.5 measurements, limiting the accuracy of PM2.5-related health studies. We develop a method to combine estimates for PM2.5 using satellite-retrieved aerosol optical depth (AOD) and simulations from the Community Multiscale Air Quality (CMAQ) modeling system. While most previous methods utilize AOD or CMAQ separately, we aim to leverage advantages offered by both methods in terms of resolution and coverage by using Bayesian model averaging. In an application of estimating daily PM2.5 in the Southeastern US, the ensemble approach outperforms statistical downscalers that use either AOD or CMAQ in cross-validation analyses. In addition to PM2.5, our approach is also highly applicable for estimating other environmental risks that utilize information from both satellite imagery and numerical model simulation.

✉ nancy.murray@emory.edu

25. GENERALIZED LINEAR MODELS

CONVERGENCE PROPERTIES OF GIBBS SAMPLERS FOR BAYESIAN PROBIT REGRESSION WITH PROPER PRIORS

Saptarshi Chakraborty*, *University of Florida*

Kshitij Khare, *University of Florida*

The Bayesian probit model (Albert and Chib (1993)) is popular and widely used for binary regression. While an improper flat prior for the regression coefficients is appropriate in the absence of prior information, a proper normal prior is desirable when prior information is available or in high dimensional settings where the no. of coefficients (p) is greater than the sample size (n). For both choices of priors, the resulting posterior density is intractable and a Data Dugmentation (DA) Markov chain is used to draw approximate samples from it. In this paper, we first show that in case of proper normal priors, the DA Markov chain is geometrically ergodic *for any* design matrix X, n and p (unlike the improper prior case, where n >= p and another condition on X are needed for posterior propriety itself). This provides theoretical guarantees for constructing standard errors for MCMC estimates. We also derive sufficient conditions under which the DA Markov chain is trace-class (i.e., the corresponding operator has summable eigenvalues). In particular, this allows us to conclude the existence of sandwich algorithms which are strictly better than the DA algorithm in an appropriate sense.

✉ c7rishi@ufl.edu

IDENTIFIABILITY AND BIAS REDUCTION IN THE SKEW-PROBIT MODEL FOR A BINARY RESPONSE

DongHyuk Lee*, *Texas A&M University*

Samiran Sinha, *Texas A&M University*

The skew-probit link function is one of the popular choices for modelling the success probability of a binary variable with regard to covariates. This link deviates from the probit link function in terms of a flexible skewness parameter. For this flexible link, the identifiability of the parameters is investigated. Next, to reduce bias of the maximum likelihood estimator of the skew-probit model we propose to use the penalized likelihood approach. We consider three different penalty functions, and compare them via extensive simulation studies. Based on the simulation results we make some practical recommendations. For the illustration purpose, we analyze a real dataset on heart-disease.

✉ dhyuklee@stat.tamu.edu

A FLEXIBLE ZERO-INFLATED COUNT MODEL TO ADDRESS DATA DISPERSION

Kimberly F. Sellers*, *Georgetown University*

Andrew Raim, *U.S. Census Bureau*

Excess zeroes are commonly associated with data overdispersion in count data, however this relationship is not guaranteed. One should instead consider a flexible distribution that not only can account for excess zeroes, but can also address potential over- or under-dispersion. We introduce a zero-inflated Conway-Maxwell-Poisson (ZICMP) regression to model the relationship between explanatory and response variables, accounting for both excess zeroes and dispersion. This talk introduces the ZICMP model and illustrates its flexibility, highlighting various statistical properties and model fit through several examples.

✉ kfs7@georgetown.edu

A ROBUST WALD TEST OF HOMOGENEITY FOR CORRELATED COUNT DATA WITH EXCESS ZEROS

Nadeesha R. Mawella*, *Kansas State University*

Wei-Wen Hsu, *Kansas State University*

David Todem, *Michigan State University*

KyungMann Kim, *University of Wisconsin, Madison*

Homogeneity tests for zero-inflated models are used to evaluate the heterogeneity in the population, where the heterogeneity often refers to the zero counts generated from two different sources. In these tests, the mixture probability that represents the extent of heterogeneity is then examined at zero. For these tests, it requires the correct model specification in the testing procedure in order to provide valid statistical inferences. However, in practice, the test could be performed with a misspecified conditional mean or an incorrect baseline distribution of the zero-inflated model, which could result in biased statistical inferences. In this paper, a robust Wald test statistic is proposed for correlated count data with excess zeros. Technically, the proposed test is developed under the framework of Poisson-Gamma model and the use of a working independence model coupled with a sandwich estimator to adjust for any misspecification of the covariance structure in data. The empirical performance of the proposed test is assessed though simulation studies. The longitudinal dental caries data from Detroit Dental Health Project is used to illustrate the proposed test.

✉ nadee@ksu.edu

GENERALIZED LINEAR MODELS WITH LINEAR CONSTRAINTS FOR MICROBIOME COMPOSITIONAL DATA

Jiarui Lu* •, *University of Pennsylvania*

Pixu Shi, *University of Pennsylvania*

Hongzhe Li, *University of Pennsylvania*

Motivated by regression analysis for microbiome compositional data, this paper considers generalized linear regression analysis with compositional covariates, where a group of linear constraints on regression coefficients are imposed to account for the compositional nature of the data and to achieve subcompositional coherence. A penalized likelihood estimation procedure using a generalized accelerated proximal gradient method is developed to efficiently estimate the regression coefficients. A de-biased procedure is developed to obtain asymptotically unbiased and normally distributed estimates, which leads to valid confidence intervals of the regression coefficients. Simulations results show the correctness of the coverage probability of the confidence intervals and smaller variances of the estimates when the appropriate linear constraints are imposed. The methods are illustrated by a microbiome study in order to identify bacterial species that are associated with inflammatory bowel disease (IBD) and to predict IBD using fecal microbiome.

✉ jiaruilu@pennmedicine.upenn.edu

A GLM-BASED LATENT VARIABLE ORDINATION METHOD FOR MICROBIOME SAMPLES

Michael B. Sohn*, *University of Pennsylvania*

Hongzhe Li, *University of Pennsylvania*

Distance-based ordination methods, such as principal coordinates analysis (PCoA), are widely used in the analysis of microbiome data. However, these methods are prone to pose a potential risk of misinterpretation about the compositional difference in samples across different populations if there is a difference in dispersion effects. Accounting for high sparsity and overdispersion of microbiome data, we propose a GLM-based Ordination Method for Microbiome Samples (GOMMS). This method uses a zero-inflated quasi-Poisson (ZIQP) latent factor model. An EM algorithm based on the quasi-likelihood is developed to estimate parameters. It performs comparatively to the distance-based approach when dispersion effects are negligible and consistently better when dispersion effects are strong, where the distance-based approach sometimes yields undesirable results.

✉ msohn@pennmedicine.upenn.edu

26. MEASUREMENT ERROR

CAUSAL INFERENCE IN THE CONTEXT OF AN ERROR PRONE EXPOSURE: AIR POLLUTION AND MORTALITY

Xiao Wu*, *Harvard School of Public Health*

Danielle Braun, *Harvard School of Public Health*

Marianthi-Anna Kioumourtzoglou, *Columbia University School of Public Health*

Christine Choirat, *Harvard School of Public Health*

Qian Di, *Harvard School of Public Health*

Francesca Dominici, *Harvard School of Public Health*

We propose a new approach for estimating causal effects when the exposure is mismeasured and confounding adjustment is performed via generalized propensity score (GPS). Using validation data, we propose a regression calibration (RC)-based correction for a continuous error-prone exposure combined with GPS to adjust for confounding after categorizing the corrected continuous exposure (RC-GPS). We consider GPS adjustment via subclassification, IPTW, and matching. In simulations, RC-GPS eliminates bias from exposure error and confounding. We applied RC-GPS to estimate the causal effect of long-term PM2.5 exposure on mortality in New England (2000-2012). The main study contains 2,202 zip codes (217,660 grids) with yearly mortality and PM2.5 averages from a spatiotemporal model (error-prone). The internal validation study includes 83 grids with error-free yearly PM2.5 averages from monitors. Under non-interference and weak unconfoundedness assumptions, we found that moderate exposure ($8<PM2.5<10\ \mu g/m^3$) causes a 2.5% (95% CI: 0.0%, 3.4%) increase in mortality compared to low exposure ($PM2.5<8\ \mu g/m^3$).

✉ wuxiao@g.harvard.edu

MEASUREMENT ERROR MODELS FOR GROUP TESTING DATA

Md S. Warasi*, *Radford University*

Joshua M. Tebbs, *University of South Carolina*

Group testing is a cost-effective procedure, where individuals are combined into pools and then pools are tested for a binary characteristic (e.g., positive or negative disease status). Recently, group testing regression models have been studied widely to make covariate-adjusted inference on individuals. However, in many applications in epidemiology and other areas, covariates cannot be measured correctly. When mismeasured covariates are used without accounting for errors, inference from group testing regression models can be highly biased. Furthermore, the problem can be aggravated when error-prone testing responses are used. In this project, we propose new regression methods for group testing data in the presence of errors in covariate measurements and testing responses. Our general approach acknowledges both types of error and offers reliable inference in the context. We illustrate our methods using simulation and real data applications.

✉ msarker@radford.edu

STATISTICAL STRATEGIES FOR THE ANALYSIS OF DIET-DISEASE MODELS THAT CORRECT FOR ERROR-PRONE EXPOSURES WITHIN A COMPLEX SURVEY DESIGN

Pedro L. Baldoni, *University of North Carolina, Chapel Hill*

Daniela T. Sotres-Alvarez*, *University of North Carolina, Chapel Hill*

Pamela A. Shaw, *University of Pennsylvania School of Medicine*

Dietary intake is typically assessed using self-reported instruments such as 24-hour dietary recalls, which are measured with systematic and random measurement error leading to biased estimators and attenuated associations. To date, there are a handful of recovery biomarkers that measure usual dietary nutrient intake and that can be

used to calibrate self-reported dietary measurements (e.g. 24-hour urinary excretion measure to calibrate sodium intake). Given its high cost and burden on the participant, recovery biomarkers are obtained in a small sample to develop calibration equations. These are used to calibrate nutrients in larger studies and to estimate their effect on health outcomes. Motivated by the Hispanic Community Health Study/Study of Latinos (HCHS/SOL) and accounting for its complex survey design, we present several statistical strategies (parametric and non-parametric bootstrap and multiple imputation) to estimate the effect of a biomarker-calibrated nutrient on health outcomes and to account for the extra variability coming from the calibration model. Simulation studies are also presented.

✉ dsotres@unc.edu

CORRECTION OF MISCLASSIFICATION ERROR IN PRESENCE OF NON-IGNORABLE MISSING DATA

Haresh Dharmu Rochani*, *Georgia Southern University*

Lili Yu, *Georgia Southern University*

Hani Samawi, *Georgia Southern University*

Missing data and misclassification errors are very common problem in many research studies. It is well known that the misclassification error in covariates can cause bias estimation of parameters for statistical model. It can also reduce the overall statistical power. Misclassification simulation extrapolation (MC-SIMEX) procedure is a well-known method to correct the bias in parameter estimation due to misclassification for given statistical model. Misclassification matrix has to be known or estimated from a validation study to use MC-SIMEX method. However, in many circumstances, the validation study has non-ignorable missing data. Estimation of misclassification matrix can be biased and hence the estimation of parameters of given statistical model in presence of non-ignorable missing data. In this paper, we apply the Baker, Rosenberger and Dersimonian modeling approach to perform the sensitivity analysis using MC-SIMEX method. Simulation studies are used to investigate the efficiency of parameters under given assumption of missing data mechanism. We illustrate the method by using "National Health and Nutrition Examination Survey" dataset.

✉ hrochani@georgiasouthern.edu

MISCLASSIFICATION SIMULATION EXTRAPOLATION PROCEDURE FOR LOG-LOGISTIC SURVIVAL DATA

Varadan Sevilimedu, *Georgia Southern University*

Lili Yu*, *Georgia Southern University*

Hani Samawi, *Georgia Southern University*

Haresh Rochani, *Georgia Southern University*

Misclassification of binary covariates is pervasive in survival data and this often leads to inaccurate parameter estimates. Despite the importance of log-logistic distribution in situations where the hazard rates do not obey monotonicity, the topic of misclassification error has not been researched in log-logistic survival data. We aim to fill the above mentioned gap in literature by developing a method involving the simulation and extrapolation algorithm, to correct for misclassification in log-logistic AFT models. The choice of this method is driven by its flexibility and minimal assumptions with survival data. Simulations are carried out with varying degrees of censoring in outcomes and misclassification in covariates. The goal is to evaluate the effectiveness of our method in correcting for the bias caused by misclassification in covariates and also to describe the impact of ignoring misclassification in log-logistic AFT models. The proposed method is applied to survival data of stage C prostate cancer patients participating in flow cytometry studies.

✉ lyu@georgiasouthern.edu

27. METHODS FOR NEXT GENERATION SEQUENCING DATA

A STATISTICAL METHOD FOR THE ANALYSIS OF MULTIPLE ChIP-Seq DATASETS

Pedro Baldoni*, *University of North Carolina, Chapel Hill*

Naim Rashid, *University of North Carolina, Chapel Hill*

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

Chromatin immunoprecipitation followed by next generation sequencing (ChIP-seq) is a technique to detect regions of protein-DNA interaction, such a transcription factor binding sites or regions containing histone modifications. A goal of the analysis of ChIP-seq data is to identify genomic loci enriched for sequencing reads pertaining to DNA bound to the factor of interest. Given the reduction of massive parallel sequencing costs, methods to detect consensus regions of enrichment across multiple samples or differential enrichment between groups of samples are of interest. Here, we present a statistical model and software to detect consensus peak regions from multiple ChIP-seq experiments through a class of Zero-Inflated Mixed Effects Hidden Markov Models. We show that the proposed model outperforms the existing methods available in the context of broad and diffuse data (H3K27me3) by accounting for the excess zeros and potential replicate-specific effects, such as sequencing depth and input control effect. The result is a novel framework that identifies broad enrichment regions across different experiments. Simulation studies and real data results will be presented.

✉ baldoni@email.unc.edu

TRENDY: SEGMENTED REGRESSION ANALYSIS OF EXPRESSION DYNAMICS FOR HIGH-THROUGHPUT ORDERED PROFILING EXPERIMENTS

Rhonda Bacher*, *University of Wisconsin, Madison*

Ning Leng, *Morgridge Institute for Research*

Li-Fang Chu, *Morgridge Institute for Research*

James A. Thomson, *Morgridge Institute for Research*

Christina Kendziorski, *University of Wisconsin, Madison*

Ron Stewart, *Morgridge Institute for Research*

High throughput expression profiling experiments with ordered conditions (e.g. time-course or spatial-course) are becoming more common for profiling detailed differentiation processes or spatial patterns. Identifying dynamic changes at both the individual gene and whole transcriptome level can provide important insights about genes, pathways, and critical time-points. We present Trendy, a freely available R package, which utilizes segmented regression models to simultaneously characterize each gene's expression pattern and summarize overall dynamic activity in ordered condition experiments. For each gene, Trendy finds the optimal segmented regression model and provides the location and direction of dynamic changes in expression. We demonstrate the utility of Trendy to provide biologically relevant results on both microarray and RNA-seq datasets.

✉ rbacher@wisc.edu

DNA COPY NUMBER VARIANTS DETECTION USING A MODIFIED INFORMATION CRITERION IN THE DNA-Seq DATA

Jaeeun Lee*, *Augusta University*

Jie Chen, *Augusta University*

DNA copy number variations (CNVs) are associated with many human diseases. Recently, CNV studies have been carried out using Next Generation Sequencing (NGS) technology that produces millions of short reads. We apply multiple change point analysis based on the 1d fused lasso regression to the CNV detection problem with NGS reads ratio data. Given the number of copy number changes, the corresponding genomic locations are estimated by fitting the 1d fused lasso. The estimated number of change points depends on which tuning parameter value is selected in the 1d fused lasso. Although several researchers have applied lasso-based methods to the CNV detection problem, they

have not yet intensely focused on how to determine the number of copy number changes. We propose a modified Bayesian information criterion, JMIC, to estimate the optimal tuning parameter in the 1d fused lasso. We show theoretically that JMIC consistently identifies the true number of change points. Our simulation studies confirm JMIC's superiority over the existing criteria. Finally, we apply the proposed method to the reads ratio data from the breast tumor HCC1952 and its matched cell line BL1954.

✉ JLEE2@augusta.edu

AN EFFECTIVE NORMALIZATION METHOD FOR METAGENOMIC COMPOSITIONAL DATA

Ruofei Du*, *University of New Mexico Comprehensive Cancer Center*

Michael Sohn, *University of Pennsylvania*

Zhide Fang, *Louisiana State University Health Sciences Center, New Orleans*

Lingling An, *University of Arizona*

Metagenomics is the genomic/genetic study of microbial samples obtained from the environment directly. To compare taxonomic abundance of microbial communities between conditions is often of primary interest. Aiming to remove the uneven library sizes, normalization is an inevitable step prior to a statistical comparative analysis. There is increasing evidence that many metagenomic sequence data may not be treated as another variant of sequence count data, especially due to its compositional characteristics. It is observed in situations a widely applied normalizations for sequence count data (e.g. RNA-Seq data) is ineffective to metagenomic compositional data. We propose a novel scaling normalization method, weighted-sum scaling (WSS) for metagenomic compositional data analysis. In addition to highly skewed compositional proportions, the overdispersion of counts, and under-sampling issues have been adequately considered. The effectiveness of the WSS method is demonstrated by its performance, in terms of statistical power and error rate controlling, compared to other normalization methods on the simulated data and real datasets with specific count sampling/shuffling steps.

✉ rdu@salud.unm.edu

FunSPU: A VERSATILE AND ADAPTIVE MULTIPLE FUNCTIONAL ANNOTATIONS-BASED ASSOCIATION TEST OF WHOLE GENOME SEQUENCING DATA

Yiding Ma*, *University of Texas Health Science Center at Houston*

Peng Wei, *University of Texas MD Anderson Cancer Center*

Despite the ongoing large-scale population-based whole genome sequencing (WGS) projects such as the NIH, NHLBI and TOPMed program, WGS-based association analysis of complex traits remains a challenge due to the large number of rare variants. External biological knowledge, such as functional annotations based on the ENCODE, may be helpful in distinguishing causal rare variants from neutral ones. However, each functional annotation can only provide certain aspect of the biological functions. Our knowledge to select the informative annotations a priori is limited while incorporating non-informative annotations will lose power. We propose a versatile and adaptive test called FunSPU (with R package) that incorporates multiple biological annotations and is adaptive at both the annotation and variant levels, thus maintaining high power even in the presence of noninformative annotations. In addition to extensive simulations, we illustrate our proposed test using the TWINSUK cohort (n=1,718) of UK10K WGS data (Walter et al, Nature 2015) based on six functional annotations: CADD, RegulomeDB, FunSeq, Funseq2, GERP++, and GenoSkyline.

✉ yma9@mdanderson.org

A POWERFUL AND DATA-ADAPTIVE TEST FOR RARE-VARIANT-BASED GxE ANALYSIS

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

Peng Wei, *University of Texas MD Anderson Cancer Center*

As 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 correctly estimating the latter under the null hypothesis of no GxE effects and the presence of neural variants. In response, we have developed a family of powerful and data-adaptive GxE tests, called "aGE" tests, in the framework of the adaptive powered score test. Using extensive simulations, we compared aGE tests with state-of-the-art rareGE and iSKAT tests. We found that aGE tests could control the Type I error rate in the presence of misspecification of the main effects, whereas rareGE and iSKAT methods suffer from inflated Type I error rate. Our tests were more resilient to inclusion of neutral variants and more powerful under a variety of scenarios. Finally, we demonstrated the performance of the proposed aGE tests using the real data. An R package named 'aGE' is available at github.

✉ tianzhong.yang@uth.tmc.edu

28. STATISTICS IN IMAGING

MATRIX DECOMPOSITION FOR MODELING MULTIPLE SCLEROSIS LESION DEVELOPMENT PROCESSES

Menghan Hu*, *Brown University*

Russell Takeshi Shinohara, *University of Pennsylvania*

Ciprian Crainiceanu, *John Hopkins University*

Ani Eloyan, *Brown University*

This project is motivated by a longitudinal magnetic resonance imaging (MRI) study of multiple sclerosis (MS) patients. The objective is to quantify the progression of MS by studying the intensity profiles of lesions appearing on the MRI scans of MS patients. Understanding the dynamic behavior of the underlying intensity trajectories is potentially useful for determining the disease stage at the time of observation, effects of treatments, and predicting outcomes at a future visit. We model the longitudinal MRI intensities as discrete observations from a functional process over time via an extension of the longitudinal functional principal component analysis model. MS patients develop lesions at seemingly random time points at various locations, hence the support of the data is unbalanced in both space and time. We define a process describing lesion incidence and develop a general statistical model for lesion development encompassing the processes observed before and after lesion incidence. To reduce the computational complexity, we use principal component bases for the functional processes.

✉ menghan_hu@brown.edu

SPARSE DYNAMIC STRUCTURAL EQUATION MODEL WITH INTEGER PROGRAMMING AND ITS APPLICATION TO LONGITUDINAL GENETIC-IMAGING DATA ANALYSIS

Nan Lin*, *University of Texas Health Science Center at Houston*

Rong Jiao, *University of Texas Health Science Center at Houston*

Momiao Xiong, *University of Texas Health Science Center at Houston*

Dynamic Bayesian network is a probabilistic framework for modeling gene regulatory networks with time-course data. A non-stationary dynamic Bayesian network assumes that the structures and parameters of the dynamic Bayesian networks vary over time. We present a novel sparse dynamic network models that is formulated as a nonsmooth optimization problem. The traditional Newton's method is an efficient tool for solving unconstrained smooth optimization problem, but is not suited for solving large nonsmooth convex problem. Proximal methods can be viewed as an extension of Newton's method. We derive proximal gradient algorithm for - penalized maximum likelihood estimation and generalized least square estimates of parameters in our model. The introduced dynamic structural equation model is coupled with integer programming algorithm. The proposed model is applied to 282 diffusion tensor images in the Alzheimer's disease Neuroimaging Initiative at five different time points. The preliminary results are encouraging.

✉ Nan.Lin@uth.tmc.edu

INCORPORATING PRIOR INFORMATION WITH FUSED SPARSE GROUP LASSO: APPLICATION TO PREDICTION OF CLINICAL MEASURES FROM NEUROIMAGES

Joanne C. Beer*, *University of Pittsburgh*

Howard J. Aizenstein, *University of Pittsburgh*

Stewart J. Anderson, *University of Pittsburgh*

Robert T. Krafty, *University of Pittsburgh*

Predicting clinical variables from whole-brain neuroimages is a high dimensional problem that requires some type of feature selection. Penalized regression is a popular embedded feature selection method for high dimensional data. For neuroimaging applications, spatial regularization using the L1 or L2 norm of the image gradient has shown good performance, yielding smooth solutions in spatially contiguous brain regions. However, correlations are likely to exist not only between neighboring voxels, but also among spatially distributed yet related groups of voxels such as those residing in the same brain networks. We propose a penalty that encourages structured, sparse, interpretable solutions by incorporating prior information about spatial and group structure among voxels. The estimator includes lasso, fused lasso, and group lasso as special cases. We present optimization steps for fused sparse group lasso penalized regression using the ADMM algorithm. Simulation studies demonstrate conditions under which fusion and group penalties together outperform either of them alone. We use fused sparse group lasso to predict continuous measures from fMRI data using the ABIDE dataset.

✉ jcb117@pitt.edu

STATISTICAL APPROACHES FOR LONGITUDINAL BRAIN LESION SEGMENTATION

Shiyu Wang*, *University of Pennsylvania*

Multiple sclerosis (MS) is a disease of the central nervous system (CNS) that is characterized by inflammation and neuroaxonal degeneration in both gray matter (GM) and white matter (WM). Magnetic resonance imaging (MRI) can be used to detect lesions in the brains of MS patients and is essential for diagnosing the disease and monitoring its progression. With longitudinal structural MRI, we can study the evolution of white matter lesions (WML) in MS patients and use statistical algorithms to measure development of WML. However, the majority of lesion segmentation algorithms focus on cross-sectional assessments of the brain and only a few concentrate on changes in WML load. Unfortunately, neither of these approaches yields a temporally consistent and accurate comprehensive longitudinal WML assessment.

To address this, we extend a fully automated segmentation algorithm that leverages intermodal coupling information to increase the accuracy of longitudinal WML delineation.

✉ vanessaswang23@gmail.com

ESTIMATING DYNAMIC CONNECTIVITY STATES IN MULTI-SUBJECT fMRI DATA

Chee-Ming Ting*, *King Abdullah University of Science and Technology*

Hernando Ombao, *King Abdullah University of Science and Technology*

Steven L. Small, *University of California, Irvine*

Jeremy I. Skipper, *University College London*

Our goal is to estimate changes in connectivity states of large-sized brain networks. Existing studies use sliding-window analysis which is unable to capture both smooth and abrupt changes simultaneously, and rely on adhoc methods for high-dimensional estimation. Another challenge is multi-subject analysis with substantial between-subjects heterogeneity in connectivity dynamics. We propose a novel approach based on Markov-switching factor model allowing dynamic connectivity states in fMRI data to be driven by few latent factors. We specify a switching vector autoregressive (SVAR) factor process to quantity time-varying directed connectivity. It enables a reliable estimation of regime change-points and massive dependencies in each regime. We develop a three-step procedure: 1) extracting factors using principal component analysis, 2) identifying connectivity regimes in a low-dimensional subspace based on the factor SVAR, 3) constructing high-dimensional connectivity metrics using subspace estimates. We explore different schemes for inference at both subject- and group-level. The method is applied to multi-subject resting-state and movie-watching fMRI data.

✉ cmting@utm.my

AVERAGING SYMMETRIC POSITIVE-DEFINITE MATRICES IN THE SPACE OF EIGEN-DECOMPOSITIONS

Brian Thomas Rooks*, *University of Pittsburgh*

Sungkyu Jung, *University of Pittsburgh*

This paper introduces a novel method for averaging in the space of symmetric positive-definite (SPD) matrices using the scaling-rotation geometric framework introduced in Jung, Schwartzman, and Groisser (2015). The sample scaling-rotation mean set is defined as the set of sample Frechet means with respect to the scaling-rotation distance. We outline an algorithm for computing candidates for the sample scaling-rotation mean, and then present conditions guaranteeing a type of uniquenes for sample scaling-rotation means, strong consistency to the population scaling-rotation mean set, and a Central Limit Theorem result. The paper concludes with applications of the sample scaling-rotation mean framework to multivariate tensor-based morphometry and diffusion tensor image data processing.

✉ btrooks88@gmail.com

29. ORAL POSTERS: NETWORK SCIENCE

29a. INVITED ORAL POSTER: THE REDUCED PC-ALGORITHM: IMPROVED CAUSAL STRUCTURE LEARNING IN LARGE RANDOM NETWORKS

Ali Shojaie*, *University of Washington*

Arjun Sondhi, *University of Washington*

We consider the task of estimating a high-dimensional directed acyclic graph, given observations from a linear structural equation model with arbitrary noise distribution. By exploiting properties of common random graphs, we develop a new algorithm that requires conditioning only on small sets of variables. The proposed algorithm, which is essentially a modified version of the PC-Algorithm, offers significant gains in both computational complexity and

estimation accuracy. In particular, it results in more efficient and accurate estimation in large networks containing hub nodes, which are common in biological systems. We prove the consistency of the proposed algorithm, and show that it also requires a less stringent faithfulness assumption than the PC-Algorithm. Simulations in low and high-dimensional settings are used to illustrate these findings. An application to gene expression data suggests that the proposed algorithm can identify a greater number of clinically relevant genes than current methods.

✉ ashojaie@uw.edu

29b. INVITED ORAL POSTER: AN INTEGRATIVE GRAPHICAL MODELING APPROACH FOR MULTIPLE HETEROGENEOUS OMICS DATA

George Michailidis*, *University of Florida*

Rapid development of high-throughput technologies has led to the collection of data across multiple molecular compartments (genomic, transcriptomic, proteomic, etc.) and across many different biological conditions. Standard graphical modeling methods are limited in their ability to deal with such heterogeneous data, which may also differ in size and noise levels, emphasizing the need for integrative approaches. In this talk, we present a statistical framework for performing two types of integrative graphical modeling. We first present an integrated approach that jointly estimates multiple Gaussian graphical models from a variety of omics sources. We next extend this approach to allow for both multiple omics sources and varying biological conditions. We discuss the computational procedure for solving the resulting optimization problems, establish theoretical properties of the proposed estimators, and illustrate their performance with an application to TCGA breast cancer data.

✉ gmichail@ufl.edu

29c. SELECTION FOR SEMIPARAMETRIC ODDS RATIO MODEL VIA ADAPTIVE SCREENING

Jinsong Chen*, *University of Illinois, Chicago*

Hua Yun Chen, *University of Illinois, Chicago*

In network detection, the semiparametric odds ratio model (ORM) has several advantages: flexible in handling both discrete and continuous data and including interactions; avoiding the problem of model incompatibility; and invariant to biased sampling design. However, for high-dimensional data, the complexity of this model induces computational burden comparing with parametric approaches, e.g., Gaussian graphical model. In this paper, we first build the partial faithfulness for network under ORM. With this theoretical justification, we then adapt two novel screening methods for ORM: two-step selection based on reduced network and step-wise selection. The theoretical supports of these methods are developed, and simulations and data application are conducted to assess the performance of our methods.

✉ jschen24@hotmail.com

29d. BayesNetBP: AN R PACKAGE FOR PROBABILISTIC REASONING IN BAYESIAN NETWORKS

Han Yu*, *State University of New York at Buffalo*

Janhavi Moharil, *State University of New York at Buffalo*

Rachael Hageman Blair, *State University of New York at Buffalo*

In this work, we present the R package, Bayes Network Belief Propagation (BayesNetBP), for the implementation of belief propagation in directed probabilistic graphical models, known as Bayesian Networks (BNs). Network inference and analysis is a popular approach to understanding complex relationships among variables. Belief propagation offers a unique layer of information that can be used to facilitate a better understanding through quantification of system-wide

changes in the network. The BayesNetBP package is the first R package to facilitate probabilistic reasoning in discrete, continuous and mixed BNs under the framework of conditional Gaussian Bayesian networks, and is independent of other commercial software. It provides novel systems-level visualizations for probabilistic reasoning in the network, and connects seamlessly with existing graphical modeling tools in R. Finally, it also provides a Shiny app that is accessible to the non-technical expert. This software supports probabilistic reasoning in any network that can be described as a directed acyclic graph, and fills a major gap in the graphical modeling tools available in R.

✉ hyu9@buffalo.edu

29e. DYNAMIC NETWORK COMMUNITY DISCOVERY

Shiwen Shen*, *University of South Carolina*

Many methods have been developed to detect the community structure of network assuming the number of communities K is known. On the other hand, with advances in computer technology, it is possible to observe the entire evolution over time of a network. In this paper, we explore methods to determine the network communities in a dynamic setting, in which not only the current, but also the historical observations of a network, are taken into consideration to produce more accurate and stable clustering results. Our work is based on the assumption of Degree Corrected Block Model (DCBM), in which degree heterogeneity is affiliated to each node. Three frameworks (PCQ, PCM, PMD) for dynamic community detection are discussed in the paper. Each of them represents a philosophical way of understanding what historical information should be borrowed to improve the current community discovery. Simulations are conducted to compare the performances of all three frameworks. In addition, guidance is provided informed by the simulation regarding the appropriate framework to use in real data.

✉ sshen@email.sc.edu

29f. ASSESSING THE EFFECTIVE DEGREE OF SNPs IN eQTL NETWORKS

Sheila Gaynor*, *Harvard University*

Maud Fagny, *Dana-Farber Cancer Institute*

John Platig, *Dana-Farber Cancer Institute*

Xihong Lin, *Harvard University*

John Quackenbush, *Dana-Farber Cancer Institute*

Network analyses are a natural approach for identifying genetic variants and genes that work together to drive disease phenotype. The relationship between SNPs and genes, captured in expression quantitative trait locus (eQTL) analysis, can be represented as a network with edges connecting SNPs and genes. Existing network methods treat such edges as fixed and known when they are most often thresholded estimates from eQTL regression. We propose a method to characterize an essential feature of nodes of eQTL networks, their centrality, that proceeds without intermediate thresholding to retain the most data on eQTLs and limit error propagation. We define the network metric of effective degree to represent how central and potentially influential a SNP is to the network, and estimate it as a function of the eQTL regressions. We apply our method to data from the GTEx project to assess whether SNPs strongly associated to particular diseases are more central to disease-specific tissues. Specifically, we use generalized additive models to assess whether SNPs associated with esophagus cancer and type 2 diabetes are more central to eQTL networks in esophageal and adipose tissue, respectively.

✉ sgaynor@fas.harvard.edu

29g. BAYESIAN INFERENCE IN NONPARANORMAL GRAPHICAL MODELS

Jami J. Mulgrave*, *North Carolina State University*

Subhashis Ghosal, *North Carolina State University*

Gaussian graphical models, where it is assumed that the variables of interest jointly follow multivariate normal distributions with sparse precision matrices, have been used to study intrinsic dependence among several variables, but the Gaussianity assumption may be restrictive in many applications. A nonparanormal graphical model is a nonparametric generalization of a Gaussian graphical model for continuous variables where it is assumed that the variables follow a Gaussian graphical model only after some unknown smooth monotone transformation. We consider a Bayesian approach in the nonparanormal graphical model using a rank likelihood which remains invariant under monotone transformations, thereby avoiding the need to put a prior on the transformation functions. On the underlying precision matrix of the transformed variables, we consider either a spike-and-slab prior or a continuous shrinkage prior on its Cholesky decomposition and use an efficient posterior Gibbs sampling scheme. We study the numerical performance of the proposed method through a simulation study and apply it on a real dataset.

WITHDRAW

✉ jnjacks3@ncsu.edu

29h. INTEGRATIVE ANALYSIS OF BRAIN FUNCTIONAL NETWORKS BASED ON ANATOMICAL KNOWLEDGE

Ixavier A. Higgins*, *Emory University*

Graphical modeling has been a powerful tool to estimate brain functional networks using fMRI data. It is also known that the brain structure often drives functional connectivity, and that the two are correlated. Although there are a slew of graphical modeling approaches to estimate the brain functional network, there has been very limited advances to estimate the brain functional connectivity while accounting for structural connectivity information. We propose a hierarchical Bayesian approach which models the functional connectivity in terms of the structural knowledge in a manner which encourages functional connections corresponding to a structural connection. It is also flexible in allowing functional connections supported by the data which lack underlying anatomical connectivity . The method is compared to existing graphical modeling approaches. Through extensive numerical studies, we demonstrate that the proposed approach performs better than existing methods which do not incorporate anatomical information. We apply the approach to the Philadelphia Neurological Cohort (PNC) data and obtain meaningful findings supported by previous evidence.

✉ ihiggin@emory.edu

30. MODERN METHODS FOR USING HISTORICAL AND OTHER AUXILIARY DATA IN ADAPTIVE CLINICAL TRIALS

DESIGN AND COST-BENEFIT ISSUES IN HISTORICAL DATA-INCORPORATING ONCOLOGY PLATFORM TRIALS

James P. Normington*, *University of Minnesota*

Somnath Sarkar, *Roche-Genentech*

Jiawen Zhu, *Roche-Genentech*

Clinical trialists are constantly searching for new ways to reduce the trial's cost and ethical risk to its enrollees. Some trial designers have suggested borrowing information from similar but already-completed clinical trials to reduce the number of patients needed for the current study. We describe a Bayesian adaptive trial designed with an industry partner of two treatments for first-line diffuse large B-cell lymphoma. Ours is a type of platform trial that uses commensurate prior methods at various interim analyses to borrow adaptively from the control group of a separate, earlier-starting

but concurrently-running trial. The design biases the trial's randomization ratio in favor of the novel treatment when the interim posterior indicates commensurability of the two control groups. We go on to speculate about conditions under which such a design is economically viable for the sponsor, given the expense involved in its maintenance. Many future compounds entering the platform may be biomarker-specific or otherwise targeted toward a relatively small subset of patients, leading to insufficient historical controls in the platform. We conclude with possible regulatory concerns.

✉ jpnormington@gmail.com

MEAN BENEFIT FOR BIOMARKER-GUIDED TREATMENT STRATEGIES

Meilin Huang*, *University of Texas MD Anderson Cancer Center*

Brian Hobbs, *University of Texas MD Anderson Cancer Center*

Precision medicine has emerged from the awareness that many human diseases are intrinsically heterogeneous with respect to their pathogenesis and composition among patients as well as dynamic over the course of therapy. Precision medicine present challenges to traditional paradigms of clinical translational, however, for which estimates of population-averaged effects from large randomized trials are used as the basis for demonstrating comparative benefit. In this article, we present a general approach for estimating the localized treatment benefit of biomarker-guided strategies when evaluated in the context of a validation design or retrospective study with adjustment for selection bias. The statistical procedure attempts to define the localized treatment benefit of a given biomarker-guided strategy for the targeted population in consideration of the treatment response surfaces, selection rule, and inter-cohort balance of prognostic determinants. Through simulation study, both reductions in bias and MSE when compared to competing methods based on generalized linear models. The methodology is also demonstrated through a proteomic study of lower grade glioma.

✉ meilin.huang.mh@gmail.com

A MULTI-SOURCE ADAPTIVE PLATFORM DESIGN FOR EMERGING INFECTIOUS DISEASES

Alexander M. Kaizer*, *Colorado School of Public Health*

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

Joseph S. Koopmeiners, *University of Minnesota*

Emerging infectious diseases challenge traditional paradigms for clinical translation of therapeutic interventions. The Ebola outbreak in West Africa was a recent example which called for alternative designs that can be sufficiently flexible to compare multiple potential treatment regimes in a context with high mortality and limited available treatment options. The PREVAIL II master protocol was designed to address these concerns by sequentially evaluating treatments in a single trial with aggressive interim monitoring to identify effective treatments as soon as possible. One shortcoming, however, is that supplemental information from controls in previous trial segments was not utilized. We address this limitation by proposing an adaptive design methodology that facilitates information sharing across possibly non-exchangeable segments using multi-source exchangeability models (MEMs). The design uses multi-source adaptive randomization to target information balance within a trial segment in relation to posterior effective sample size. Compared to the standard design, we demonstrate that MEMs with adaptive randomization can improve power with limited type-I error inflation.

✉ alex.kaizer@ucdenver.edu

31. STATISTICAL ADVANCES IN HEALTH POLICY RESEARCH

ROBUST ESTIMATION FOR MULTIPLE UNORDERED TREATMENTS

Sherri Rose*, *Harvard Medical School*

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

Postmarket comparative effectiveness and safety analyses of therapeutic treatments typically involve large observational cohorts. We propose robust machine learning estimation techniques for implantable medical device evaluations where there are more than two unordered treatments. We isolate the effects of individual drug-eluting stents on a composite outcome. This flexible approach accommodates a large number of covariates from clinical databases while also accounting for clustering by hospital. Data from the Massachusetts Data Analysis Center (Mass-DAC) percutaneous coronary intervention cohort is used to assess the composite outcome of 10 drug-eluting stents among adults implanted with at least one drug-eluting stent in Massachusetts. We find remarkable discrimination between stents.

✉ rose@hcp.med.harvard.edu

CAUSAL APPROACHES TO COST AND COST-EFFECTIVENESS ANALYSIS WITH TIME-DEPENDENT TREATMENT REGIMES

Andrew J. Spieker*, *University of Pennsylvania*

Jason A. Roy, *University of Pennsylvania*

Nandita Mitra, *University of Pennsylvania*

Studies seeking to compare cost and cost-effectiveness across treatments are often complicated by censoring, whereby complete cost data are only available on a subset of participants. Inverse weighting approaches have previously been developed for estimation of intent-to-treat effects to address this challenge. We propose a nested g-computation approach to cost analysis appropriate for estimation of joint causal effects—i.e., contrasts in marginal mean costs under different hypothetical treatment regimes. Joint causal effects are often better able to provide insights into health policy recommendations and resource allocation. This approach naturally lends itself to cost-effectiveness analyses involving net monetary benefit. The acceptability curve has been proposed as a way of graphically summarizing net monetary benefit over a range of willingness-to-pay thresholds, though this lacks a meaningful clinical interpretation. We propose an approach based on potential outcomes that we refer to as the cost effectiveness determination curve. We apply these methods to endometrial cancer patients from SEER-Medicare data and compare results to those of existing approaches.

✉ aspieker@upenn.edu

OPTIMAL MATCHING APPROACHES IN HEALTH POLICY EVALUATIONS UNDER ROLLING ENROLLMENT

Lauren Vollmer*, *Mathematica Policy Research*

Jiaqi Li, *Mathematica Policy Research*

Jonathan Gellar, *Mathematica Policy Research*

Bonnie Harvey, *ComScore*

Sam Pimentel, *University of California, Berkeley*

In many policy evaluations, we construct a matched control group similar to the treatment group to minimize selection bias. Rolling enrollment in treatment introduces several complications. In this talk, we discuss two such complications. The first concerns the definition of the baseline period, typically 12 to 24 months pre- enrollment, used to define key matching covariates such as baseline health services utilization and expenditures. Control subjects never enroll and thus have no an enrollment date, so we must define their baseline period differently. The second complication is a common feature of rolling enrollment studies; an acute event, such as a stroke, often triggers enrollment for the treatment group. Failure to account for this acute event induces substantial selection bias. We discuss several

strategies to handle these complications, including a novel optimal matching approach that forbids a unique potential control to match to more than one treatment subject. We also develop an R implementation that is compatible with popular matching packages such as optmatch and MatchIt. Lastly, we apply our proposed method to ongoing national long-term care policy evaluations.

✉ lvollmer@mathematica-mpr.com

32. INTEGRATIVE ANALYSIS OF MULTI-OMICS DATA WITH APPLICATIONS TO PRECISION MEDICINE

STATISTICAL AND INFORMATIC ISSUES IN INTEGRATING GENOMIC AND IMAGING DATA

Debashis Ghosh*, *Colorado School of Public Health*

With the advent of 'big-data' biological sources, the array of questions that scientists and clinicians can answer has substantially expanded. Increasingly, there is simultaneous consideration of genomic and imaging data that is being considered in various medical fields, such as cancer, Alzheimer's disease and Parkinson's disease. In this talk, we will describe data acquistion and processing issues in some of these types of studies. One type of methodology that has received much attention with a single big-data source is kernel machines, and in this talk, we will describe the development of kernel machines to these settings.

✉ debashis.ghosh@ucdenver.edu

BAYESIAN VARIABLE SELECTION FOR MULTI-LAYER OVERLAPPING GROUP STRUCTURE IN LINEAR REGRESSION AND CLUSTERING SETTINGS WITH APPLICATIONS TO MULTI-LEVEL OMICS DATA INTEGRATION

George Tseng*, *University of Pittsburgh*

Li Zhu, *University of Pittsburgh*

Variable selection is a pervasive question in modern high-dimensional data analysis where the number of features often exceeds the sample size. Incorporation of group structure prior knowledge to improve variable selection has been widely developed. In this paper, we consider prior knowledge of a multi-layer overlapping group structure to improve variable selection in regression setting and clustering setting. In genomic applications, for instance, a biological pathway contains tens to hundreds of genes and a gene can contain multiple experimentally measured features (such as its mRNA expression, copy number variation and possibly methylation level of multiple sites). In addition to the hierarchical structure, the groups may be overlapped (e.g. two pathways often contain same genes). We propose a Bayesian hierarchical indicator model that can conveniently incorporate the multi-layer overlapping group structure in variable selection for regression and clustering settings. The results not only enhance prediction accuracy but also improve variable selection and model interpretation.

✉ ctseng@pitt.edu

PATHWAY-AND NETWORK-BASED INTEGRATIVE BAYESIAN MODELING OF MULTIPLATFORM GENOMICS DATA

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

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

Min Jin Ha, *University of Texas MD Anderson Cancer Center*

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

Elizabeth J. McGuffey, *United States Naval Academy*

The identification of gene pathways and networks involved in cancer development and progression and characterization of their activity in terms of multiplatform genomics can provide information leading to discovery of new targeted medications. We propose a two-step model that integrates multiple genomic platforms, as well as gene pathway membership information, to efficiently and simultaneously (a) identify the

genes significantly related to a clinical outcome, (b) identify the genomic platform(s) regulating each important gene, and (c) rank the pathways by importance to clinical outcome. We propose hierarchical Bayesian pathway- and network-based frameworks, which 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. The approaches will be illustrated using several case examples integrating high-throughput pan-omic (e.g. genomic, epigenomic, transcriptomic, proteomic) across multiple tumor types.

✉ veera@mdanderson.org

INTEGRATING LARGE-SCALE SEQUENCING DATA FOR CANCER CLASSIFICATION

Ronglai Shen*, *Memorial Sloan-Kettering Cancer Center*

In this talk, I will present a pan-cancer analysis of multiple omic data platforms from the Cancer Genome Atlas. A kernel regression approach was developed to systematically integrate sequencing data sets for predicting patient survival outcome across 14 cancer types in over 3,000 tumor samples. In addition, I will discuss the prognostic relevance of tumor sequencing in a clinical setting. We performed a mutational analysis of >300 cancer-associated genes in 1,054 patients with metastatic lung adenocarcinoma sequenced prospectively. A statistical learning framework was developed to stratify patients into different risk groups with regard to overall survival.

✉ shenr@mskcc.org

33. FUNCTIONAL DATA ANALYSIS IN BIOSCIENCES

SCALAR-ON-IMAGE REGRESSION VIA THE SOFT-THRESHOLDED GAUSSIAN PROCESS

Ana-Maria Staicu*, *North Carolina State University*

Jian Kang, *University of Michigan*

Brian J. Reich, *North Carolina State University*

We study spatial variable selection for scalar-on-image regression. We propose a new class of Bayesian nonparametric models, soft-thresholded Gaussian processes and develop the efficient posterior computation algorithms. Theoretically, soft-thresholded Gaussian processes provide large prior support for the spatially varying coefficients that enjoy piecewise smoothness, sparsity and continuity, characterizing the important features of imaging data. Also, under some mild regularity conditions, the soft-thresholded Gaussian process leads to the posterior consistency for both parameter estimation and variable selection for scalar-on-image regression, even when the number of true predictors is larger than the sample size. The proposed method is illustrated via simulations, compared numerically with existing alternatives and applied to Electroencephalography (EEG) study of alcoholism.

✉ astaicu@ncsu.edu

FUNCTIONAL DATA ANALYSIS WITH HIGHLY IRREGULAR DESIGNS WITH APPLICATIONS TO HEAD CIRCUMFERENCE GROWTH

Matthew Reimherr*, *The Pennsylvania State University*

Justin Petrovich, *The Pennsylvania State University*

Carrie Daymont, *The Pennsylvania State Health Milton S. Hershey Medical Center*

Functional Data Analysis often falls into one of two branches, either sparse or dense, depending on the sampling frequency of the underlying curves. However, methods for sparse FDA often still rely on having a growing number of observations per subject as the sample size grows. Practically, this means that for very large sample sizes with infrequently or irregularly sampled curves, common methods may still suffer a non-negligible bias. This becomes especially true for nonlinear models, which are often defined based on complete curves. In this talk I will discuss how this issue can be fixed to obtain valid statistical inference regardless of the sampling frequency of the curves. This work is motivated by a study by Dr. Carrie Daymont from Hershey medical school that examines pathologies related

to head circumference growth in children. In her study, tens of thousands of children are sampled, but with widely varying frequency.

✉ mreimherr@psu.edu

OUTLIER DETECTION IN DYNAMIC FUNCTIONAL MODELS

Andrada E. Ivanescu*, *Montclair State University*

Ciprian M. Crainiceanu, *Johns Hopkins University*

William Checkley, *Johns Hopkins University*

We present methods for dynamic identification of outliers in a longitudinal data context. We call these methods dynamic because the associated models can be applied at and tailored to any particular point in the history of the data for one individual. Dynamic approaches are different from static approaches that use all available data. Static approaches are useful in retrospective studies when one is interested in data quality control, whereas dynamic approaches are useful when one is interested in identifying unusual observations as soon as possible and use these findings for interventions as data are acquired. The methods we propose can use covariate adjustment both for time-dependent and independent covariates. Methods are motivated by and applied to a child growth study conducted in Lima, Peru.

✉ ivanescua@montclair.edu

MATRIX FACTORIZATION APPROACHES TO ANALYSIS OF FUNCTIONAL COUNT DATA

Jeff Goldsmith*, *Columbia University*

Daniel Backenroth, *Columbia University*

Jennifer Schrack, *Johns Hopkins University*

Taki Shinohara, *University of Pennsylvania*

We present a novel decomposition of non-negative functional count data, which we call NNFPCA (non-negative functional principal components analysis), that draws on ideas from non-negative matrix factorization. Our decomposition enables the study of patterns of variation across subjects in a highly interpretable manner. FPCs are estimated directly on the data scale, are local and represent `parts' that are transparently combined together via addition. This contrasts with generalized FPC approaches, which estimate FPCs on a latent scale, where decompositions of observations reflect highly complex patterns of cancellation and multiplication of FPCs that often vary across their entire domain. We apply our decomposition method to a dataset comprising observations of physical activity for elderly healthy Americans.

✉ jeff.goldsmith@columbia.edu

34. NEW ADVANCES IN ANALYSIS OF SURVIVAL DATA FROM BIASED SAMPLING

SEMIPARAMETRIC MODEL AND INFERENCE FOR BIVARIATE SURVIVAL DATA SUBJECT TO BIASED SAMPLING

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

Jin Piao, *University of Southern California*

Yu Shen, *University of Texas MD Anderson Cancer Center*

Despite treatment advances in breast cancer, some patients will have cancer recurrence within 5 years of their initial treatments. To better understand the relationship between patient characteristics and their survival after an intermediate event such as the local and regional cancer recurrence, it is of interest to analyze ordered bivariate survival data. A registry database reflects the real-world patient population, and provides a valuable resource for investigating these associations. One challenge in analyzing registry data is that the observed bivariate times tend to be longer than those in the target population due to the sampling scheme. We propose to jointly model the ordered bivariate survival data using a copula model and appropriately adjusting for

the sampling bias. We develop an estimating procedure to simultaneously estimate the parameters for the marginal survival functions and the association parameter in the copula model.

✉ jning@mdanderson.org

EFFICIENT SECONDARY ANALYSIS IN TWO PHASE STUDIES

Haibo Zhou*, *University of North Carolina, Chapel Hill*

Yinghao Pan, *Fred Hutchinson Cancer Research Center*

Two-phase sampling design has been widely used to reduce the cost for studies with time to event outcome. In the real studies, it is seldom that there is only one endpoint of interest. Investigators often would like to re-use the existing data to study the association between the exposure variables and a secondary endpoint. This is referred as secondary analysis. In this talk, we propose a restricted maximum likelihood estimator based on the empirical likelihood corresponding to the two-phase sampling design. We jointly model the time-to-event outcome and the outcome of interest in secondary analysis. The advantage of our method is that it is efficient and yet require no strong parametric assumptions on the covariate distributions.

✉ zhou@bios.unc.edu

ESTIMATION OF GENERALIZED SEMIPARAMETRIC REGRESSION MODELS FOR THE CUMULATIVE INCIDENCE FUNCTIONS WITH MISSING COVARIATES

Yanqing Sun*, *University of North Carolina, Charlotte*

Unkyung Lee, *Texas A&M University*

Thomas H. Scheike, *University of Copenhagen*

Peter B. Gilbert, *University of Washington and Fred Hutchinson Cancer Research Center*

The cumulative incidence function quantifies the percentage of failures over time due to a specific cause for competing risks data. We investigate the generalized semiparametric regression models for the cumulative incidence functions when covariates may have missing values from a two-phase sampling design or by missing-at-random. The effects of some covariates are modeled as nonparametric functions of time while others are modeled as parametric functions of time. Different link functions can be selected to add flexibility in modeling the cumulative incidence functions. We develop estimation procedures based on the direct binomial regression and the inverse probability weighting of complete cases. The approaches for improving estimation efficiency are also investigated. The asymptotic properties of the proposed estimators are established. The simulation studies show that the proposed estimators have satisfactory finite-sample performances. The methods are applied to analyze data from the RV144 vaccine efficacy trial to investigate the associations of immune response biomarkers with the cumulative incidence of HIV-1 infection.

✉ yasun@uncc.edu

FITTING ACCELERATED FAILURE TIME MODEL USING CALIBRATED WEIGHTS FOR CASE-COHORT STUDIES

Sangwook Kang*, *Yonsei University*

Dahhay Lee, *Yonsei University*

A case-cohort design is an efficient study design for analyzing failure time data by reducing the cost and effort of conducting a large cohort study. Estimation of regression coefficients is typically done through a weighted estimating equation approach whose weight is the inverse of the sampling probabilities. Several techniques to enhance the efficiency by estimating weights or calibrating weights based on auxiliary variables have been developed for Cox models. In this paper, we propose to extend these methodologies to semiparametric accelerated failure time models.

✉ kanggi1@yonsei.ac.kr

35. ESTIMATION AND OPTIMIZATION FOR THE EFFECTS OF SCREENING SCHEDULES AND TREATMENT TIMING

DESIGNING SCREENING TESTS THAT MINIMIZE THE TIME BETWEEN INFECTION AND POSITIVE DIAGNOSIS IN HIV/AIDS

Robert Strawderman*, *University of Rochester*

John Rice, *University of Colorado, Denver*

Brent Johnson, *University of Rochester*

We consider the possibility of designing screening regimens that minimize the time between infection and positive diagnosis (i.e. through testing) in HIV/AIDS and other diseases having comparatively low infection rates. Assuming that testing occurs according to an underlying renewal process, we use tools from renewal theory to derive approximations to the expected value of this time frame. It is shown, in particular, that the best testing regimen is that which minimizes the coefficient of variation of the inter-test times. Simulation studies are used to study this phenomenon and the impact of non-regular testing procedures on the increase in the expected time between infection and diagnosis.

✉ robert_strawderman@urmc.rochester.edu

MODELING THE EFFECT OF CANCER SCREENING ON MORTALITY

Alex Tsodikov*, *University of Michigan*

Mortality represents a time-to-event endpoint expressed by the age at cancer-specific death. Primary treatment is applied at an intermediate event represented by cancer diagnosis. The timing of primary treatment is affected by cancer screening. More intensive screening leads to generally earlier diagnosis and treatment for the disease. Assessing the effect of a screening strategy on mortality is a causal modeling exercise, because treatment is confounded by indication, and because the timing of treatment in this context is always informative of mortality, even if it is ineffective. We use semiparametric joint and mechanistic models to study the question.

✉ tsodikov@umich.edu

CAUSALITY IN THE JOINT ANALYSIS OF LONGITUDINAL AND SURVIVAL DATA

Lei Liu*, *Washington University in St. Louis*

Cheng Zheng, *University of Wisconsin, Milwaukee*

Joseph Kang, *Centers for Disease Control and Prevention*

In many biomedical studies, disease progress is monitored by a biomarker over time, e.g., repeated measures of CD4, hemoglobin level in end stage renal disease (ESRD) patients. The endpoint of interest, e.g., death or diagnosis of a specific disease, is correlated with the longitudinal biomarker. The causal relation between the longitudinal and time to event data is of interest. In this paper we examine the causality in the analysis of longitudinal and survival data. We consider four questions: (1) whether the longitudinal biomarker is a mediator between treatment and survival outcome; (2) whether the biomarker is a surrogate marker; (3) whether the relation between biomarker and survival outcome is purely due to an unknown confounder; (4) whether there is a mediator moderator for treatment. We illustrate our methods by data from two clinical trials: an AIDS study and a liver cirrhosis study.

✉ lei.liu@wustl.edu

OPTIMAL TIMING OF STEM CELL TRANSPLANT FOR LEUKEMIA PATIENTS

Xuelin Huang*, *University of Texas MD Anderson Cancer Center*

Xiao Lin, *University of Texas MD Anderson Cancer Center*

Jorge Cortes, *University of Texas MD Anderson Cancer Center*

Patients with chronic myeloid leukemia may go through a chronic phase, accelerated phase and blast crisis. While they are in early chronic phase, there are a few targeted therapies that can bring their disease under control without causing severe toxicities. At this time, when the risk of death is low, stem cell transplant, which is potentially dangerous, may not be a good option. However, by the time of blast crisis, when the risk of death is very high, it might be too late to receive a transplant. Then a natural question is that, during this process, when is the best time for transplant? We answer this question by analyzing the patients diagnosed after the year 2001. Some of them never received a transplant, others received transplant during different disease stages, with different types of donors, including siblings, well-matched or partially-matched non-siblings, and other types. The identification of the optimal timing, accounting for patient status and the availability of donor types, is important for guiding clinical practice. Various statistical methods are used for these data analyses, including multistate Markov model and dynamic prediction models.

✉ xlhuang@mdanderson.org

36. CLUSTERED DATA METHODS

ROBUST CLUSTERING WITH SUBPOPULATION-SPECIFIC DEVIATIONS

Briana Joy K. Stephenson* •, *University of North Carolina, Chapel Hill*

Amy H. Herring, *Duke University*

Andrew Olshan, *University of North Carolina, Chapel Hill*

The National Birth Defects Prevention Study (NBDPS) is a case-control study of birth defect etiology conducted across 10 US states. Researchers are interested in characterizing the etiologic role of maternal diet on the development of congenital malformations, using data tools such as the food frequency questionnaire. In a large, heterogeneous population, traditional clustering methods, such as latent class analysis, used to estimate dietary patterns can produce a large number of clusters due to a variety of factors, including study size and regional diversity. These factors result in a loss of interpretability that may differ due to minor consumption pattern changes. Motivated by the local partition process, we propose a new method, Robust Profile Clustering, where participants may cluster at two levels: (1) globally, where women are assigned to an overall population-level cluster, and (2) locally, where variations in diet are accommodated via a Beta-Bernoulli process dependent on subpopulation differences. Using NBDPS data, we use our method to derive dietary patterns of pregnant women in the US while accounting for regional variability.

✉ bjks@live.unc.edu

AN IMPROVED DISSIMILARITY MEASURE FOR CLUSTERING DATA WITH MIXED DATA TYPES

Shu Wang*, *University of Pittsburgh*

Jonathan G. Yabes, *University of Pittsburgh*

Chung-Chou H. Chang, *University of Pittsburgh*

Discovering patterns in the data has never been more relevant in the era of Big Data. One of the most widely used approaches is cluster analysis. Dissimilarity metrics, by which most clustering algorithms rely on, are well developed for continuous variables. However, almost all clinical datasets contain categorical variables. For such mixed data types, Gower's Distance is the popular dissimilarity metric but using it leads to clustering results that categorical variables dominated. Hence, we propose a dissimilarity measure that can overcome this limitation by using proportion of a variable's between-cluster sum of dissimilarity in total dissimilarity. In simulated datasets that contain mixed data types, our proposed measure could achieve higher adjusted rand index and that the variables' weights reflect clustering contribution appropriately. We applied our proposed method to identify possible sepsis phenotypes among all sepsis patients admitted to eight intensive care units of a hospital during eight calendar periods. Variables used were abstracted from hospital electronic health records that included patient information at the time of admission.

✉ shw97@pitt.edu

CLUSTER-STRATIFIED OUTCOME-DEPENDENT SAMPLING IN RESOURCE-LIMITED SETTINGS: INFERENCE AND DESIGN CONSIDERATIONS

Sara Sauer*, *Harvard School of Public Health*

Sebastien Haneuse, *Harvard School of Public Health*

Bethany Hedt-Gauthier, *Harvard Medical School*

Catherine Kirk, *Partners in Health Rwanda*

Alphonse Nshimyiryo, *Partners in Health Rwanda*

Faced with limited resources, public health program evaluations often resort to using routinely collected group-level data rather than the patient-level data needed to answer nuanced evaluation questions. Cluster-stratified sampling, in which clinics are sampled and detailed information on all patients in the selected clinics then collected, offers a cost-efficient solution. Given data from a cluster-stratified design, Cai et al. (2001) proposed estimation for a marginal model using inverse-probability-weighted generalized estimating equations. Towards performing inference, however, the variance expression presented by Cai et al. (2001) ignored covariance in the cluster-specific selection indicators. We provide a corrected variance expression, as well as a consistent plug-in estimator. Simulations are conducted to examine the small-sample operating characteristics of the proposed method, and to explore the potential efficiency gains of sampling designs in which readily-available group-level information guides cluster selection. The proposed methods are illustrated using birth data from 18 clinics in Rwanda, collected via a cluster-stratified scheme.

✉ ssauer@g.harvard.edu

PAIRWISE COVARIATES-ADJUSTED BLOCK MODEL FOR COMMUNITY DETECTION

Sihan Huang*, *Columbia University*

Yang Feng, *Columbia University*

The stochastic block model (SBM) is one widely used community detection model for network data. However, SBM is restricted by the strong assumption that all nodes in the same community are stochastically equivalent, which may not be suitable for practical applications. We introduce pairwise covariates-adjusted stochastic block model (PCABM), a generalization of SBM that incorporates pairwise covariate information. In our model, the pairwise covariates can be constructed using any bivariate function of the corresponding covariates of the pair of nodes considered. We study the maximum likelihood estimators of the coefficients for the covariates as well as the community assignments and show they are consistent under typical sparsity conditions. Spectral clustering with adjustment (SCWA) is introduced to efficiently solve PCABM. Under certain conditions, we derive the error bound of community estimation under SCWA and show that it is community detection consistency. PCABM compares favorably with the SBM or degree-corrected stochastic block model (DCBM) under a wide range of simulated and real networks when covariate information is accessible.

✉ sh3453@columbia.edu

A WEIBULL-COUNT APPROACH FOR HANDLING UNDER- AND/OR OVER-DISPERSED CLUSTERED DATA STRUCTURES

Martial Luyts*, *I-Biostat and Katholieke Universiteit Leuven*

Geert Molenberghs*, *I-BioStat, Hasselt University and Katholieke Universiteit Leuven*

Geert Verbeke, *I-Biostat and Katholieke Universiteit Leuven*

Koen Matthijs, *Katholieke Universiteit Leuven*

Clarice Demétrio, *University of São Paulo, Brazil*

John Hinde, *NUI Galway, Ireland*

A Weibull-model-based approach is examined to handle under- and/or over-dispersed count data in a hierarhical framework. This methodology was first introduced by Nakagawa et al. (1975), and later examined for under- and

over-dispersion by Kalktawi et al. (2015) in the univariate case. Extensions to hierarchical approaches with under- and over-dispersion were left unnoted, even though it can be obtained in a simple manner. Here, a random-effects extension of the Weibull-count model is proposed and compared with some well-known models in the literature. The empirical results indicate that this approach is a useful general framework in the context of clustered discrete data.

✉ martial.luyts@kuleuven.be

HOMOGENEITY TEST OF RISK RATIOS FOR STRATIFIED CORRELATED BINARY DATA

Yuqing Xue*, *State University of New York at Buffalo*

Chang-Xing Ma, *State University of New York at Buffalo*

In stratified ophthalmologic (or otolaryngologic) studies, correlated bilateral data often arise along with potential confounding effect when information of paired body parts are collected from each individual across strata. In this article, we investigate three testing procedures for testing the homogeneity of the relative risk across strata with correlated binary data. Our simulation results indicate that the score testing procedure usually produces relatively satisfactory type I error control with reasonable power, hence is recommended. We further illustrate our proposed methods with a randomized trial example from an ophthalmologic study.

✉ yuqingxu@buffalo.edu

SAMPLE SIZE DETERMINATION FOR GEE ANALYSES OF STEPPED WEDGE CLUSTER RANDOMIZED TRIALS

Fan Li* •, *Duke University*

Elizabeth L. Turner, *Duke University*

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

In stepped wedge cluster randomized trials, clusters of individuals switch from control to intervention from a randomly-assigned period onwards. Such trials are becoming increasingly popular in health service research. When a closed cohort is recruited from each cluster for longitudinal follow-up, proper sample size calculation should account for three distinct types of correlations: the within-period, the inter-period and the within-individual correlations. Setting the latter two correlations to be equal accommodates cross-sectional designs. We propose sample size procedures for continuous and binary responses within the framework of generalized estimating equations that employ a block exchangeable within-cluster correlation structure defined from the distinct correlation types. For continuous responses, we show that the intraclass correlations affect power through two eigenvalues of the correlation matrix. We demonstrate that analytical power agrees well with the simulated power even for as few as 8 clusters, when data are analyzed using bias-corrected estimating equations for the correlation parameters concurrently with a bias-corrected sandwich variance estimator.

✉ frank.li@duke.edu

37. ADVANCES IN CAUSAL INFERENCE

ASSESSING SENSITIVITY TO UNMEASURED CONFOUNDING WITH MULTIPLE TREATMENTS AND BINARY OUTCOMES: A BAYESIAN APPROACH

Liangyuan Hu*, *Icahn School of Medicine at Mount Sinai*

Chenyang Gu, *Harvard University*

Michael Lopez, *Skidmore College*

Unmeasured confounding in observational studies gives rise to biased treatment effect estimates, and sensitivity analysis (SA) is a solution to assessing the magnitude of these biases. There are no known techniques for assessing unmeasured confounding with multiple treatments and binary outcomes. We propose a flexible Bayesian SA framework for unmeasured confounding under the multiple treatment setting. We first derive the biases in treatment effect estimates when the assumption of no unmeasured confounding is violated, with the magnitude and direction of the violation governed by sensitivity parameters. We then

develop a flexible semi-parametric Bayesian approach utilizing BART, that corrects the bias attributable to unmeasured confounding. Inferences about the sensitivity to unmeasured confounding will be obtained from the posterior distribution of the average treatment effects. Extensive simulations are carried out to investigate the joint impact of sensitivity parameters and various combinations of design factors for the multiple treatment setting, including confounding level, response surfaces and treatment assignment mechanism.

✉ liangyuan.hu@mountsinai.org

CAUSAL INFERENCE FOR INTERFERING UNITS FOR CLUSTER AND POPULATION LEVEL INTERVENTION PROGRAMS

Georgia Papadogeogrou*, *Harvard University*

Fabrizia Mealli, *University of Florence*

Corwin Zigler, *Harvard School of Public Health*

Interference arises when an individual's potential outcome depends on the individual treatment level, but also on the treatment level of others. A common assumption in the causal inference literature in the presence of interference is partial interference, implying that the population can be partitioned in clusters of individuals whose potential outcomes only depend on the treatment of units within the same cluster. Previous literature has defined average potential outcomes under counterfactual scenarios where treatments are randomly allocated to units within a cluster. However, within clusters there may be units that are more or less likely to receive treatment based on covariates or neighbors' treatment. We define estimands that describe average potential outcomes for realistic regimes taking into consideration the units' covariates, as well as dependence between units' treatment assignment. We discuss these estimands, propose unbiased estimators and derive asymptotic results as the number of clusters grows. Finally, we estimate effects in a comparative effectiveness study of emission reduction technologies on ambient ozone concentrations in the presence of interference.

✉ gpapadogeorgou@fas.harvard.edu

A BAYESIAN REGULARIZED MEDIATION ANALYSIS WITH MULTIPLE EXPOSURES

Yu-Bo Wang*, *Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health*

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

Jill M. Goldstein, *Harvard Medical School, Brigham and Women's Hospital and Massachusetts General Hospital*

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

Stephen E. Gilman, *Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health*

Mediation analysis assesses the effect of study exposure on an outcome both through and around specific mediators. While mediation analysis involving multiple mediators has been addressed in recent literature, the case of multiple exposures has received little attention. With the presence of multiple exposures, we consider regularizations that allow simultaneous effects selection and estimation, while stabilizing model fit and accounting for model selection uncertainty. In the framework of linear structural-equation models, we show that a two-stage approach regularizing regression coefficients does not guarantee a unimodality and that a product-of-coefficient approach regularizing effects of interest tends to over penalize. We propose a regularized difference-of-coefficient approach that bypasses these limitations. Using the connection between regularization and Bayesian hierarchical models, we develop an efficient Markov chain Monte Carlo algorithm for posterior estimation and inference. Through simulations, we show that the proposed approach has better empirical performances compared to some alternatives. The methodology is illustrated using two epidemiological studies.

✉ yu-bo.wang@nih.gov

AN ALTERNATIVE ROBUST ESTIMATOR OF AVERAGE TREATMENT EFFECT IN CAUSAL INFERENCE

Jianxuan Liu*, *Bowling Green State University*

Yanyuan Ma, *The Pennsylvania State University*

Lan Wang, *University of Minnesota*

The problem of estimating average treatment effect is important when evaluating the effectiveness of medical treatments or social intervention policies. Most of the existing methods for estimating average treatment effect rely on some parametric assumptions on the propensity score model or outcome regression model one way or the other. We propose an alternative robust approach to estimating the average treatment effect based on observational data in the challenging situation when neither a plausible parametric outcome model nor a reliable parametric propensity score model is available. Our approach has the advantage of being robust, flexible, data adaptive and it can handle many covariates simultaneously. Adapting a dimension reduction approach, we estimate the propensity score weights semiparametrically by using a nonparametric link function to relate the treatment assignment indicator to a low-dimensional structure of the covariates. We demonstrate the robust performance of the estimators on simulated data and a real data example of analyzing the effect of maternal smoking on babies' birth weight.

✉ jianxuanliu7@gmail.com

THE GENERALIZED FRONT-DOOR FORMULA FOR ESTIMATION OF INDIRECT CAUSAL EFFECTS OF A CONFOUNDED TREATMENT

Isabel R. Fulcher*, *Harvard University*

Ilya Shpitser, *Johns Hopkins University*

Eric Tchetgen Tchetgen, *Harvard University*

The Population Intervention Effect of an exposure measures the expected change of an outcome from its observed value, if one were to withhold the exposure in the entire population. This effect is of interest in settings that evaluate the impact of eliminating a harmful exposure from a population. This paper develops methodology to identify and estimate the extent to which the exposure affects the outcome through an intermediate variable in a setting where the exposure may be subject to unmeasured confounding. The identifying formula for this Population Intervention Indirect Effect (PIIE) is shown to amount to a generalization of Pearlâ's front-door formula. Although Pearl's front-door recovers the indirect effect when exposure is confounded, it relies on the stringent assumption of no direct effect of exposure on outcome. The generalized front-door is shown to apply whether or not such direct effect is present. Parametric and semiparametric estimators of the PIIE are proposed and evaluated in a simulation study. Finally, the methods are applied to measure the effectiveness of monetary saving recommendations for delivery among pregnant women enrolled in a health program in Zanzibar.

✉ isabelfulcher@g.harvard.edu

LONGITUDINAL VARIABLE SELECTION IN CAUSAL INFERENCE WITH COLLABORATIVE TARGETED MINIMUM LOSS-BASED ESTIMATION

Mireille E. Schnitzer*, *Université de Montréal*

Joel Sango, *Statistics Canada*

Steve Ferreira-Guerra, *Université de Montréal*

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

Causal inference methods have been developed for longitudinal observational study designs where confounding is thought to occur over time. In particular, marginal structural models model the expectation of the counterfactual outcome conditional only on past treatment and possibly a set of baseline covariates. In such contexts, model covariates are generally identified using domain-specific knowledge. However, this may leave an analyst with a large set of potential confounders that may hinder estimation. Previous approaches to data-adaptive variable selection in causal inference focused on the single time-point setting. We

develop a longitudinal extension of collaborative targeted minimum loss-based estimation (C-TMLE) for the estimation of the parameters in a marginal structural model that can be applied to perform variable selection in propensity score models. We demonstrate the properties of this estimator through a simulation study and apply the method to investigate the safety of trimester-specific exposure to inhaled corticosteroids during pregnancy in women with mild asthma.

✉ mireille.schnitzer@umontreal.ca

38. IMPUTATION APPROACHES WITH MISSING DATA

BAYESIAN REGRESSION ANALYSIS FOR HANDLING COVARIATES WITH MISSING VALUES BELOW THE LIMIT OF DETECTION

Xiaoyan Lin*, *University of South Carolina*

Haiying Chen, *Wake Forest School of Medicine*

Environmental and biomedical research often produces data below the limit of detection(LOD). When the value is below LOD, it is denoted as non-observable. These non-observable values can be truly 0 or some undetectable nonzero values. Simply treating non-observable values as a fixed value between 0 and LOD or ignoring the 0 component to impute data has been shown to produce biased inferences in general. In this talk, we investigate the effect on the regression analysis when using different methods to impute the non-observable covariate values. We compare the Bayes MCMC imputation method with the traditional multiple imputation methods by simulations and a real data application.

✉ lin9@mailbox.sc.edu

ANALYSIS OF BINARY RESPONSE ENDPOINT WITH MISSING DATA IN SMALL STUDY POPULATION

JD Lu*, *Bioverativ, Inc.*

In evaluating response rate for small patient population, any missing data can considerably influence the study outcome. Despite diligent efforts to minimize the missing data, patients may withdraw from the treatment, receive rescue treatment, or simply miss visit due to logistic reasons. Non-response imputation is less satisfactory in drawing an inference about the response rate, especially in the rare disease setting. This abstract introduces a method of "exact" imputation in small sample size studies, where the probability of any potential outcome is derived based on the assumption of missing at random. The estimates for the response rate and its treatment difference can then be calculated by incorporating the probabilities, and its statistical inference (e.g., confidence intervals) will be drawn via bootstrapping. The author intends to compare the proposed method with other methods such as GEE, multiple imputation, and bound estimates by Horowitz and Manski.

✉ jdlu2013@gmail.com

MAXIMUM LIKELIHOOD ESTIMATION IN REGRESSION MODELS WITH CENSORED COVARIATES

Jingyao Hou*, *University of Massachusetts, Amherst*

Jing Qian, *University of Massachusetts, Amherst*

The problem of censored covariates arises frequently in family history studies, in which an outcome of interest is regressed on an age of onset, as well as in longitudinal cohort studies, in which biomarkers may be measured post-baseline. Use of censored covariates without any adjustment is well known to lead to bias in estimates of the coefficients of interest and inflated type I error. We propose an expectation maximization (EM) algorithm for estimations based on full likelihoods involving infinite-dimensional parameters under regression models with randomly censored covariates. The procedure allows joint estimation of

regression coefficients and the distribution function of the random censored covariate. Estimation procedures under both linear and generalized linear models are developed. Simulation studies show that the proposed methods perform well with moderate sample size and lead to more efficient estimators compared to complete-case analysis. We illustrate the proposed methods in application to an Alzheimer's disease study.

✉ jingyaohou@gmail.com

MULTIPLE IMPUTATION USING BOOTSTRAP

Hejian Sang*, *Iowa State University*

JaeKwang Kim, *Iowa State University*

Multiple imputation (MI) is widely used to handle missing data problem. Current MI method is developed and established under Bayesian framework, which requires intensive computation, if the posterior does not have explicit distribution. Furthermore, due to congeniality assumption, the imputation model is very sensitive and important. However, model selection and evaluation for the imputed model is not established. In this paper, we propose a new MI method using Bootstrap. The proposed method is calibrated to MI under the Bayesian framework. The proposed MI can incorporate the frequentist model selection method simultaneously and reflect the model uncertainty under the finite sample. The asymptotic properties are established. Limited simulation studies are presented to validate the proposed MI and compare with MI under the Bayesian framework.

✉ hjsang@iastate.edu

SEQUENTIAL REGRESSION IMPUTATION IN MULTILEVEL DATA

Gang Liu*, *State University of New York at Albany*

Recai Yucel, *State University of New York at Albany*

To overcome the problem of item nonresponse with skip patterns, bounds, and diverse measurement in the analysis of correlated survey data, fully-parametric multiple imputation (MI) inference can offer a viable solution. Recently, Yucel and colleagues (Yucel et al. 2017) proposed a sequential hierarchical regression imputation which is tailored to impute the correlated missing values with diverse measurement. Their algorithm employed computational techniques based on Markov Chain Monte Carlo (MCMC) and/or numerical integration are applied to approximate the conditional posterior predictive distributions. We extend these methods to allow higher levels of observational units. In particular, we consider a three level of nesting and present our computational algorithm. A comprehensive simulation study assessing the key operational characteristics as well as compatibility of this approach with the joint data generation mechanism is also discussed.

✉ lg.statistics@gmail.com

IMPUTATION FOR INVESTIGATION OF SOCIAL SUPPORT AND DISABILITY PROGRESSION IN MULTIPLE SCLEROSIS

Anastasia M. Hartzes*, *University of Alabama at Birmingham*

Stacey S. Cofield, *University of Alabama at Birmingham*

Strong social support is associated with positive outcomes for long-term disease. Consequently, destabilization of relationships can negatively influence long-term outcomes and quality of life. Consistent data-capture over time can pose statistical problems in assessing the relationship between social support and disease outcomes. Using semi-annual marital status and disease severity surveys from 2007-12 from the North American Research Committee on Multiple Sclerosis (NARCOMS), marital status was imputed for those with 2 or fewer consecutive missing responses, using a modified last observation carried forward approach to determine effects of missing information of disease outcome. Of 4474 who met inclusion criteria, 42.0% (1908) had complete data for the 5 year study period (12 update surveys); number missing

ranged 0-6 surveys per person, for 7 missingness patterns. 4703 responses in 2566 persons were imputed. Without imputation at 10 years, 72% were still married/cohabitating, with imputation 69%. Imputation was similar but underestimated the percent with social support. Imputation should be approached with caution, justified by statistical and disease considerations.

✉ ahartzes@uab.edu

39. METHODS FOR LONGITUDINAL DATA ANALYSIS

ANALYSIS OF LONGITUDINAL DATA WITH OMITTED ASYNCHRONOUS LONGITUDINAL COVARIATES

Li Chen*, *University of Missouri*

Hongyuan Cao, *University of Missouri*

Long term follow-up with longitudinal data is common in many medical investigations. In such studies, some longitudinal covariates can be omitted for various reasons. In cross sectional studies, coefficient estimation of a covariate is unbiased if the covariate is orthogonal to the omitted covariate. This is not true in longitudinal data analysis, where omission of time dependent covariate can lead to biased coefficient estimate even if the corresponding covariate is orthogonal to the omitted longitudinal covariate. In this article, we propose a new unbiased estimation method to accommodate omitted longitudinal covariate. In addition, if the omitted longitudinal covariate is asynchronous with the longitudinal response, we propose a two stage approach for valid statistical inference. Asymptotic properties of the proposed parameter estimates are established. Extensive simulation studies provide numerical support for the theoretical findings. We illustrate the performance of our method on a dataset from an HIV study.

✉ lichen@mail.missouri.edu

ANALYSIS OF THE HIGH SCHOOL LONGITUDINAL STUDY DATA TO ASSESS THE ROLES OF MENTORSHIP ON MATHEMATICS ACHIEVEMENT AND STUDENTS' INTENTIONS TO ENROLL IN STEM PROGRAMS

Anarina L. Murillo*, *University of Alabama at Birmingham*

Olivia Affuso, *University of Alabama at Birmingham*

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

Many efforts have aimed to increase student recruitment into science, technology, engineering, and mathematical (STEM) careers. However, despite many efforts student recruitment and retention in STEM majors remain low in the United States, and is much lower among underrepresented minorities. It is well-understood that mentors play a significant role in fostering students; academic development, which can be enhanced through student engagement in STEM programs. Here we utilize the High School Longitudinal Study (2009-2013) dataset to evaluate the factors associated with students; intention to pursue a STEM major which includes: mathematics achievement, mentorship, and student participation in STEM activities. Hence, the aim of this work is to assess the significance of these stated factors in order to give insight into STEM education policy efforts. Our hope is that this work would shed light on the roles of mentors and motivate future programs, workshops, and extracurricular activities to recruit and retain students in STEM majors, and particularly in biostatistics, bioinformatics, computer science, and the mathematical sciences.

✉ amurillo@uab.edu

A COMPARISON STUDY OF METHODS FOR LONGITUDINAL BIOMARKER COMBINATION: HOW RANDOM OBSERVATION TIME PLAYS A ROLE

Yongli Justin Han*, *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*

In studying prognostic biomarkers, an important topic is the prediction of a binary endpoint based on longitudinal biomarker data. An example is the Scandinavian Fetal Growth Study, which concerns predicting poor pregnancy outcomes, such as macrosomia, by utilizing ultrasound measurements collected during fetal development. Depending on the study design, the biomarkers may be measured at different time points for different individuals. However, it is not well understood how the random observation time may affect the estimation of biomarker combination, and the subsequent risk prediction. In our work, we compare the performance of several recently developed methods, shared random effect method, pattern mixture model, sufficient dimension reduction method, and some machine learning methods, due to the fact that this is a supervised learning problem. Under the designs of fixed and random observation times, prediction accuracy and risk calibration characteristics are compared in extensive simulation studies and the fetal growth data. Practical recommendations on applications of the above methods are further made.

✉ yongli.han@nih.gov

REGRESSION MODELING OF LOGIC RULES FOR LONGITUDINAL DATA

Tan Li*, *Florida International University*

Wensong Wu, *Florida International University*

Ingrid Gonzalez, *Florida International University*

In many researches, interaction effects are essential to discover associations or make predictions. In most of regression methodologies, it is hard to find the effect of complex interaction but only simple interactions (two-way or three-way). However, the complex interaction between more than three predictors may have strong effect on response, especially when all the predictors are binary. Boolean logic expression is one way to express different complex interactions between multiple binary predictors with "not", "and", and "or" logic relations. This situation, for instance, arises when a disease diagnostic is desired based on a group of binary symptoms. It is needed to develop a regression methodology to discover and evaluate the strength of such Boolean logic expressions in regression modeling. A lot of times, longitudinal data is collected in public health, which may induce correlations among observations. The purpose of this paper is going to develop regression modeling of logic rules such as Boolean logic statement for longitudinal data. The performance will be evaluated in simulation studies and real data analysis.

✉ tanli@fiu.edu

NON-GAUSSIAN LONGITUDINAL DATA ANALYSIS

Ozgur Asar*, *Acibadem University*

David Bolin, *Chalmers University of Technology*

Peter Diggle, *Lancaster University*

Jonas Wallin, *Lund University*

WITHDRAW

In this study, we consider linear mixed effects models with non-Gaussian random components for analysis of longitudinal data with large number of repeats. The modelling framework postulates that observed outcomes can be de-composed into fixed effects, subject-specific random effects, a continuous-time stochastic process, and random noise. Likelihood-based inference is implemented by a computationally efficient stochastic gradient algorithm. Random components are predicted by either of filtering or smoothing distributions. The R package ngme provides functions to implement the methodology.

✉ ozgurasarstat@gmail.com

AN R2 STATISTIC FOR COVARIANCE MODEL SELECTION IN THE LINEAR MIXED MODEL

Byron Casey Jaeger* •, *University of Alabama at Birmingham*

Lloyd J. Edwards, *University of Alabama at Birmingham*

The linear mixed model (LMM), sometimes referred to as the multi-level model, stands as one of the most widely used tools for analyses involving clustered data. Various definitions of R2 have been proposed for the LMM, but several limitations prevail. Presently there is no definition of R2 in the LMM that accommodates (1) an interpretation based on variance partitioning, (2) a method to quantify uncertainty and produce confidence limits, and (3) a capacity to conduct covariance model selection in a manner similar to information criteria. In this article, we introduce an R 2 for the LMM with each of these three characteristics. The proposed R2 measures the proportion of generalized variance explained by fixed predictors in the model. Illustrative longitudinal data from a sleep deprivation study are used to demonstrate an application of the proposed R2.

✉ bcjaeger@uab.edu

40. MICROBIOME RESEARCH METHODS

A SPARSE REGRESSION FRAMEWORK FOR INTEGRATING PHYLOGENETIC TREE IN PREDICTIVE MODELING OF MICROBIOME DATA

Li Chen*, *Auburn University*

Jun Chen, *Mayo Clinic*

The development of next generation sequencing offers an opportunity to predict the disease outcomes of patients using the microbiome sequencing data. Considering a typical microbiome dataset consists of more taxa than samples, and all the taxa are related to each other is the phylogenetic tree, we propose a smoothness penalty-Laplacian penalty to incorporate the prior information of phylogenetic tree to achieve coefficient smoothing in a sparse regression model. Moreover, we observe that sparsifying the Laplacian matrix usually results better prediction performance, however, the optimal threshold for sparsifying varies dataset by dataset and is unknown in real data analysis. To overcome this limitation, we further develop another phylogeny-constraint penalty based on evolutionary theory to smooth the coefficients with respect to the phylogenetic tree. Using simulated and real datasets, we demonstrate that the proposed methods has better prediction performance than the other competing methods.

✉ li.chen@auburn.edu

A PERMUTATION FRAMEWORK FOR DIFFERENTIAL ABUNDANCE ANALYSIS OF MICROBIOME SEQUENCING DATA

Jun Chen*, *Mayo Clinic*

One central theme of microbiome studies is to identify differentially abundant bacteria associated with some outcome. Many methods have been proposed for this task. Among these, count-based parametric models are especially appealing due to good interpretability and high statistical power. However, the excessive zeros in the microbiome sequencing data make the traditional asymptotic theory-based inference for count model not robust. To address the limitations, we propose a permutation framework for differential abundance analysis. The framework is based on weighted least squares estimation for linear models and hence is computationally efficient. It takes into account the full characteristics of microbiome data including variable library sizes, the correlations among taxa, the compositional nature of the sequencing data, and the phylogenetic relatedness of the taxa. By simulations, we show that the proposed method is more robust than current parametric methods without sacrifice of much statistical power. We also demonstrate the power of our method by applying to several real data sets.

✉ chen.jun2@mayo.edu

SPARSE HIGH-DIMENSIONAL PRECISION MATRIX ESTIMATION FOR COMPOSITIONAL DATA

Rong Ma*, *University of Pennsylvania*

Yuanpei Cao, *University of Pennsylvania*

Hongzhe Li, *University of Pennsylvania*

Motivated by the problem of estimating the microbial networks in microbiome studies, this paper proposes an estimator of sparse high-dimensional precision matrix for compositional data. Due to the simplex constraint of compositional data, the covariance matrix of the underlying basis counts is not identifiable. However, the covariance matrix based on the centered log-ratio transformation, which is identifiable, can be a good approximation to the basis covariance matrix under high-dimensional and sparse setting. The proposed estimator of the precision matrix takes advantage of this approximation and is constructed by solving a constrained L1 minimization problem. Rates of convergence are derived under spectral norm, sup-norm and Frobenius norm. A stability selection algorithm is proposed for support and sign recovery. The method is applied to UK Twins gut microbiome data set to obtain microbial interaction network.

✉ rongm@mail.med.upenn.edu

PERFect: PERMUTATION FILTRATION OF MICROBIOME DATA

Ekaterina Smirnova*, *University of Montana*

Snehalata Huzurbazar, *West Virginia University*

Farhad Jafari, *University of Wyoming*

Microbiota composition, which is associated with a number of diseases including obesity and bacterial vaginosis, requires preprocessing steps that take into account sparsity of counts and large number of taxa. Filtering is defined as removing taxa that are present in a small number of samples and have small counts in the samples where they are observed. Currently, there is no consensus on filtering standards and quality assessment. This can adversely affect downstream analyses and reproducibility of results. We introduce PERFect (https://github.com/katiasmirn/PERFect), a novel filtering approach designed to address two unsolved problems in microbiome data processing: (i) define and quantify loss due to filtering, and (ii) introduce and evaluate a permutation test for filtering loss to provide a measure of excessive filtering. Methods are assessed on mock data, where the true taxa compositions are known, and are applied to a vaginal microbiome data set. PERFect correctly removes contaminant taxa, quantifies filtering loss, and provides a uniform data-driven filtering criteria for real microbiome data sets.

✉ ekaterina.smirnova@mso.umt.edu

4I. PERSONALIZED / PRECISION MEDICINE

RELATIVE EFFICIENCY OF PRECISION MEDICINE DESIGNS FOR CLINICAL TRIALS WITH PREDICTIVE BIOMARKERS

Weichung Joe Shih*, *Rutgers University*

Yong Lin, *Rutgers University*

Prospective randomized clinical trials addressing biomarkers are time-consuming and costly, but are necessary for regulatory agencies to approve new therapies with predictive biomarkers. For this reason, recently there have been many discussions and proposals of various trial designs and comparisons of their efficiency in the literature. We compare statistical efficiencies between the marker-stratified design and the marker-based precision medicine design regarding testing/estimating four hypotheses/parameters of clinical interest, namely: treatment effects ineach marker positive and negative cohorts, marker-by-treatment interaction, and the marker's clinical utility. We quantify the relative efficiency as a function of design factors including the marker-positive prevalence rate, marker assay and classification sensitivity and specificity, and the treatment randomization ratio. It is interesting to examine the trends of the relative efficiency with these design parameters in

testing different hypotheses. We advocate to use the stratified design over the precision medicine design in clinical trials with predictive biomarkers.

✉ w.joe.shih@rutgers.edu

PartialCOXEN: IN VITRO GENE EXPRESSION-BASED PREDICTION OF RESPONSE TO ANTICANCER DRUG IN CANCER PATIENTS

Youngchul Kim*, *Moffitt Cancer Center and Research Institute*

Several sophisticated methods were introduced to develop multi-gene expression predictor for response of cancer patients to anticancer drugs based on gene expression and pharmacologic data of large-scale cancer cell line panels. However, it still remains challenging to address accurate predictions because of biological discrepancy between cancer cell lines and patient tumors. Furthermore, those methods often ignored their clinic-pathologic features that can affect drug responses. We thus propose non-interaction regression analysis (NIRA) and partial co-expression extrapolation (PartialCOXEN) algorithm to incorporate those features into identifying drug sensitivity genes with concordant co-expression across cancer cell lines and patient tumors. Drug response predictors were then trained on expression data of the genes in cancer cell lines. We applied NIRA and PartialCOXEN to the Cancer Cell Line Encyclopedia data to develop a gene expression predictor for paclitaxel response in ovarian cancer patients. A 53-genes expression model achieved significantly accurate predictions in three ovarian cancer cohorts and outperformed predictors built by other conventional approaches.

✉ youngchul.kim@moffitt.org

PREDICTION AND PREVENTION OF HOSPITAL ADVERSE EVENTS USING ROUTINELY COLLECTED PATIENT DATA WITH A MECHANISM FOR EVALUATING EFFECTIVENESS

Henry J. Domenico*, *Vanderbilt University Medical Center*

Daniel W. Byrne, *Vanderbilt University Medical Center*

Preventable harm occurring during the course of health care is responsible for an estimated 400,000 deaths/yr. Models for identifying patients at risk of these events have been published, however, these often suffer from limitations that prevent them from being used in the clinical workflow. In addition, lack of evaluation prevents effective prevention strategies from being distinguished from those that are ineffective. To overcome these obstacles, we have identified the key features that prediction models must have to be adopted into the workflow and developed models that meet these criteria. We have also developed a method for displaying patient risk within an electronic health record (EHR) and evaluating the effectiveness of prevention strategies. Data on 104620 encounters over a three year period at a large academic medical center were collected using the EHR. Regression was used to predict patient level risk of multiple clinically important adverse events. Development of models that meet the criteria for clinical usefulness can be developed and integrated into the workflow. Prevention strategies centered around these models can be evaluated using a randomized design.

✉ henry.domenico@vanderbilt.edu

A STOCHASTIC SEARCH APPROACH TO STUDY HETEROGENEITY OF TREATMENT EFFECT

Yang Hu*, *Beth Israel Deaconess Medical Center, Harvard Medical School*

Changyu Shen, *Beth Israel Deaconess Medical Center, Harvard Medical School*

Existing statistical methods to identify sub-groups with differential treatment benefit/harm are either based on some parametric structure of the underlying data generation mechanism and/or are estimated through local optimization. We developed a nonparametric approach to identify subgroups through global optimization. Our approach is composed of two steps. In the first step, a discretization procedure creates a number of small sub-populations called "cells" with sufficient granularity, which serves as the building blocks of subgroup identification. In the second step, a simulated annealing algorithm is used to search for combinations of the cells that yield up to three groups: those

deriving benefit from the treatment, those harmed by the treatment and the rest. Simulation studies are performed to evaluate the performance of this algorithm as compared with existing methods. A real data example is also presented.

✉ yhu2@bidmc.harvard.edu

ASSESSING SNP EFFECTS ON TREATMENT EFFICACY IN TAILORED DRUG DEVELOPMENT: ISSUES AND REMEDIES

Yue Wei*, *University of Pittsburgh*

Ying Ding, *University of Pittsburgh*

There has been increasing interest in discovering personalized medicine in current pharmaceutical drug development and medical research using SNPs. Finding SNPs that are predictive of treatment efficacy, measured by a clinical outcome, is fundamentally different from association detection for a quantitative trait. Some common practices often start with testing for complete null within each SNP and then use the p-values to rank all the SNPs. In personalized medicine, clinical effect size matters, and an important decision to make is to identify which genetic subgroup(s) of patients should be the target of a drug, instead of to discover which genetic characteristics are associated with the disease. In this research, we discuss the issues with some common practices and provide potential remedies. Specifically, for each SNP, we provide simultaneous confidence intervals directed toward detecting possible dominant, recessive, or additive effects. Across the SNPs, we control the expected number of SNPs with at least one false confidence interval coverage. This approach provides a step toward confidently targeting a patient subgroup in a tailored drug development process.

✉ yuw95@pitt.edu

SPARSE CONCORDANCE-ASSISTED LEARNING FOR OPTIMAL TREATMENT DECISION

Shuhan Liang*, *North Carolina State University*

Wenbin Lu, *North Carolina State University*

Rui Song, *North Carolina State University*

Lan Wang, *North Carolina State University*

To find optimal decision rule, Fan et al. (2016) proposed an innovative concordance-assisted learning algorithm which is based on maximum rank correlation estimator. It makes better use of the available information through pairwise comparison. However the objective function is discontinuous and computationally hard to optimize. In this paper, we consider a convex surrogate loss function to solve this problem. In addition, our algorithm ensures sparsity of decision rule and renders easy interpretation. We derive the L2 error bound of the estimated coefficients under ultra-high dimension. Simulation results of various settings and application to STAR*D both illustrate that the proposed method can still estimate optimal treatment regime successfully when the number of covariates is large.

✉ sliang4@ncsu.edu

42. ORAL POSTERS: HEALTH SERVICES AND HEALTH POLICY

42a. INVITED ORAL POSTER: PROFILING MEDICAL PROVIDERS USING METHODS BASED ON THE EMPIRICAL NULL

John D. Kalbfleisch*, *University of Michigan*

Lu Xia, *University of Michigan*

Zhi Keving He, *University of Michigan*

Yanming Li, *University of Michigan*

It is important to monitor patient outcomes of health care providers in order to identify problems as they arise with the general aim of improving the quality of care. In this presentation, we consider methods that account for the natural, unexplained variation among providers using methods based on the empirical null, whereby the majority of providers is used to create a national standard or norm to which all providers are compared. This approach takes account of the

number of patients served by the provider. The methods are extended to allow for a proportion of the random provider effects to be partitioned into two independent components, one part due to quality of care and one part due to incomplete risk adjustment. In this, a specified proportion of the underlying between provider variation is assumed to be due to variation in the quality of care. This gives a continuum of approaches, all based on the use of fixed effects estimates of the provider effects but with varying assessments of the allowable variation. Empirical Bayes and shrinkage estimates are also discussed, and the methods are illustrated using data on readmission rates of dialysis patients.

✉ jdkalbfl@umich.edu

42b. INVITED ORAL POSTER: POINTWISE MUTUAL INFORMATION AND SIMILARITY INDICES TO IDENTIFY TREATMENTS AND DIAGNOSES IN SEER-MEDICARE

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

Tian Bai*, Temple University*

Ashis Chanda*, Temple University*

Richard J. Bleicher, *Fox Chase Cancer Center*

Slobodan Vucetic*, Temple University*

Linked Surveillance Epidemiology and End Results (SEER)-Medicare data is used for research. Identification of treatments and comorbidities in claims requires aggregating clusters of ICD and CPT codes. Often, expert knowledge is used to identify relevant codes. However, expert identification can be difficult due to the number of codes and temporal changes. We are developing algorithms and software to better ensure that SEER-Medicare research is more reliable when using billing codes. Specifically, we have modified a Pointwise Mutual Information statistic (PMI). Heuristically, PMI is created by taking the log of the ratio of the observed probability of two codes co-occurring divided by the probability that the two codes would occur under independence. Larger PMI values indicate that two codes are more likely to occur together. We have also developed a similarity index to determine how often two ICD-9 or CPT codes cluster together, either directly or indirectly through third codes. The similarity index is created in part by factorizing our modified PMI value matrix. Values closer to one indicate more complete similarity of two codes. We developed software for these methods.

✉ Brian.Egleston@fccc.edu

42c. INVITED ORAL POSTER: INTRODUCTION TO TOOLS FOR LEARNING AND IMPLEMENTING BAYESIAN ADAPTIVE DESIGNS

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

Compared to the frequentist method, Bayesian approach offers many advantages as it is more intuitive, directly addressing the question of interest, properly accounting for uncertainty, allowing flexible and frequent trial monitoring, naturally incorporating prior information, permitting the construction of utility function for decision making with multiple outcomes (e.g., efficacy and toxicity, cost and effectiveness), etc. Bayesian methods take the "learn as we go" approach and are innately suitable for clinical trials. Many innovative Bayesian adaptive designs have been proposed to identify better treatments in a timely, efficient, accurate, and cost-effective way. However, a big challenge is how to communicate the Bayesian thinking to researchers who may not be well versed in statistics. In addition, there are relatively few tools for learning Bayesian update, gauging the impact of the prior distribution, designing and implementing Bayesian trials. At University of Texas MD Anderson Cancer Center, we have developed many downloadable and online tools for learning and implementing Bayesian adaptive designs. For example, computational and visualization tools are available for the Bayesian analysis based on beta-binomial distribution, normal-normal distribution, normal-inverse gamma distribution, diagnostic test, ROC curve analysis, etc. The impact of the prior distribution

can be easily displayed for the sensitivity analysis. Bayesian adaptive designs and analysis tools are available for dose finding, posterior and predictive probability calculations, outcome adaptive randomization, multi-arm platform design, multi-endpoint design, and hierarchical modeling, etc. Bayesian adaptive clinical trial designs increase the study efficiency, allow more flexible trial conduct, and treat more patients with more effective treatments in the trial but also possess desirable frequentist properties. The operating characteristics of various designs can be evaluated via simulations. Easy-to-use programs are available for the design and implementing of Bayesian adaptive designs. Some of these useful software's will be demonstrated. All are freely available at the followings two sites: https://biostatistics.mdanderson.org/softwareOnline/ and https://biostatistics.mdanderson.org/softwareDownload/.

✉ jjlee@mdanderson.org

42d. BAYESIAN HIERARCHICAL MULTIVARIATE POISSON REGRESSION MODELS FOR CHARACTERIZING THE DIFFUSION OF NEW ANTIPSYCHOTIC DRUGS

Chenyang Gu*, *Harvard Medical School*

Haiden Huskamp, *Harvard Medical School*

Julie Donohue, *University of Pittsburgh*

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

New treatment technologies are the primary driver of spending growth in the United States with physicians playing a pivotal role in their adoption. Studies of physician prescribing behavior indicate that the placement of a particular physician on the adoption curve for one drug does not necessarily predict where that physician falls for other drugs. We propose a new model to summarize the diffusion paths across therapeutically similar antipsychotics, in which the diffusion path of each drug is modeled by a semiparametric Poisson model with physician-specific random effects. The joint model is constructed by concatenating the univariate models based on a correlated random effects assumption. We use Markov Chain Monte Carlo methods to obtain posterior inferences. We propose performance indices to identify fast adopters of antipsychotics based on posterior tail probabilities of relevant model parameters and determine which set of physicians' covariates are related to the adoption patterns. Methods are illustrated by using dispensing information for 16932 physicians between Jan 1,1997 and Dec 31,2007 from the Xponent database, maintained by Quintiles IMS.

✉ gu@hcp.med.harvard.edu

42e. CHALLENGES OF USING ELECTRONIC HEALTH RECORDS FOR RISK PREDICTION IN ADULTS WITH TYPE 2 DIABETES

Douglas David Gunzler*, *Case Western Reserve University*

Risk prediction methods can be used to identify members of subgroups of adults with type 2 diabetes (DM2) who have similar risk levels for adverse DM2 health outcomes, independent of any treatment exposure, and have a sufficient likelihood of benefitting from treatment. Electronic health record systems registries can be used for extracting data on a sample of DM2 adults from a large, diverse hospital population for retrospective analyses. Risk measures based on central tendency measures assume that all individuals in a sample have similar exposure to risk factors and disease progression over a life course. Older DM2 individuals often have better DM2 control than younger individuals, despite the progressive nature of the disease. We discuss methods to identify risk groups of DM2 individuals that respond differently to exposure to risk factors for complications and mortality; members of these risk groups face different mortality pressures as they age. Treatment and care approaches can be assigned to potential responders with better precision according to group risk level.

✉ dgunzler@metrohealth.org

42f. ROBUST INTERRUPTED TIME SERIES MODEL FOR ASSESSING AN INTERVENTION IN MULTIPLE HEALTHCARE UNITS

Maricela F. Cruz*, *University of California, Irvine*

Miriam Bender, *University of California, Irvine*

Daniel L. Gillen, *University of California, Irvine*

Hernando Ombao, *King Abdullah University of Science and Technology*

Care delivery is complex with interacting and interdependent components that challenge traditional statistical analytic techniques, especially when modeling a time series "interrupted" by a change in health care delivery. Interrupted time series (ITS) is a robust quasi-experimental design able to infer the effectiveness of an intervention while accounting for data dependency. We develop a statistical model, 'Robust Multiple-ITS', that allows for the estimation of a common change point across hospital units in the presence of a lagged treatment effect. Thus, Robust Multiple-ITS estimates (rather than assume) the over-all time delay between formal intervention implementation and the intervention's effect on the outcomes of interest. We conduct simulations to determine the sample size required to estimate the change point under various magnitudes of pre-/post-intervention effects, based on true change point coverage probabilities. The Robust Multiple-ITS model is illustrated by analyzing staff productive hours and patient satisfaction data from a hospital that implemented and evaluated a new nursing care delivery model in multiple patient care units.

✉ maricelafcruz55@gmail.com

42g. PROPENSITY SCORE MATCHING FOR MULTILEVEL SPATIAL DATA: ACCOUNTING FOR GEOGRAPHIC CONFOUNDING IN HEALTH DISPARITY STUDIES

Melanie L. Davis*, *Medical University of South Carolina*

Brian Neelon, *Medical University of South Carolina*

Paul J. Nietert, *Medical University of South Carolina*

Kelly J. Hunt, *Medical University of South Carolina*

Lane F. Burgette, *RAND Corporation*

Andrew B. Lawson, *Medical University of South Carolina*

Leonard E. Egede, *Medical College of Wisconsin*

To explore racial disparities in diabetes specialty care among veterans, we introduce a spatial propensity score matching method to account for "geographic confounding", which occurs when confounding factors vary by geographic region. We augment models with spatial random effects, which are assigned conditionally autoregressive priors to improve inferences by borrowing information across neighboring regions. In simulation we show that ignoring spatial heterogeneity results in increased absolute bias while incorporating spatial random effects yields improvement. In the application, we construct multiple estimates of the risk difference in diabetes care: an unadjusted estimate, an estimate based on patient-level matching, and an estimate that further incorporates spatial information. The unadjusted estimate suggests that specialty care is more prevalent among non-Hispanic blacks, while patient-level matching indicates that it is less prevalent. Hierarchical spatial matching supports the latter conclusion, with a further increase in the magnitude of the disparity. These results suggest the need for culturally sensitive and racially inclusive clinical care.

✉ davml@musc.edu

42h. SHORTENING PATIENT REPORTED OUTCOME MEASURES WITH OPTIMAL TEST ASSEMBLY

Daphna Harel*, *New York University*

Patient-reported outcome measures (PROs) - such as aspects of mental health - assess aspects of patients' lives from their own perspective. Patients enrolled in clinical trials or observational studies may be asked to respond to many different scales to provide information regarding their experiences or treatment response. Efficient measurement

of PROs is thus essential to limit patient burden and research cost. However, methods to shorten these instruments are under-developed, leading to several shortened versions of the same PRO. Optimal test assembly (OTA) is an application of linear programming used frequently for item selection in designing high-stakes educational tests that incorporates the results of an IRT model to select a subset of an item pool that best satisfies pre-specified constraints while optimally maximizing an objective function, such as total test information. This presentation shows how OTA may be used to shorten PROs in an objective, reproducible, and replicable way to produce optimal shortened forms, and compares the use of OTA to the selection of items based on factor loadings. The utility of this method then applied to the PHQ-9.

✉ daphna.harel@nyu.edu

42i. TRENDS IN TRACT-LEVEL OBESITY RATES IN PHILADELPHIA BY RACE, SPACE, AND TIME

Yaxin Wu*, *Drexel University*

Dina Terloyeva, *Drexel University*

Harrison Quick, *Drexel University*

While recent data indicate that Philadelphia County has among the highest rates of adult obesity among the ten most highly populated counties in the United States, other research indicates significant disparities both spatially and by race. The goal of this work is to investigate both racial and geographic disparities in obesity rates in Philadelphia Census tracts over the period 2000 – 2015. Our data consist of self-reported survey responses from Public Health Management Corporation's Community Health Data Base's Southeastern Pennsylvania Household Health Survey. To analyze these data – and to obtain more reliable rate estimates – we apply the multivariate space-time conditional autoregressive model, simultaneously accounting for spatial-, temporal-, and between-race dependence structures. By doing so, we are able to observe temporal trends in and make inference on geographic- and racial disparities in tract-level obesity rates in Philadelphia.

✉ yw574@drexel.edu

42j. BIOSTATISTICIANS IN THE INDUSTRY HAVE REAL WORLD IMPACT ON THE OPIOID CRISIS

Meridith Blevins Peratikos*, *Axial Healthcare, Inc.*

Opioid overdoses have quadrupled since 1999 with over 33,000 deaths in 2015. On August 10, 2017, the U.S. President declared the opioid crisis a national emergency. And with nearly half of overdose deaths involving a prescription opioid, the healthcare industry must play a role in resolving this national crisis. This presentation will highlight the real-world impact a biostatistician may contribute towards combating a national emergency. A simple algorithm was developed to increase the effectiveness of practitioner outreach for opioid prescribing practices. The initial step was to gather project requirements from key stakeholders, including the pharmacists who make phone calls and the software engineers who develop the outreach portal. Data were wrangled together from disparate sources, such as contact data from the outreach portal, and administrative health claims from a SQL database. The goal was to assemble a "minimum viable" product in 3 weeks. The algorithm was demonstrated using data visualization in order to clearly communicate the proposed change to business leaders for adoption. Early results will be shared on the real-world impact of this simple algorithm.

✉ mperatikos@axialhealthcare.com

43. RECENT INNOVATIONS IN PRACTICAL CLINICAL TRIAL DESIGN

UTILITY-BASED DESIGNS FOR CLINICAL TRIALS WITH MULTIPLE OUTCOMES

Thomas A. Murray*, *University of Minnesota*

Ying Yuan, *University of Texas MD Anderson Cancer Center*

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

Utility-based methods are presented for the design and monitoring of randomized comparative clinical trials in chronic lymphocytic leukemia and surgical oncology. Numerical utilities of all elementary events are elicited to quantify their desirabilities. These numerical values are used to reduce the outcome probabilities for each treatment to a mean utility. The mean utilities reflect the clinical desirability of the average outcome for each treatment and are used as a one-dimensional criterion for monitoring the trial. Bayesian group sequential tests are discussed for two probability models, the Dirichlet-multinomial model and a novel cumulative logistic regression model.

✉ murra484@umn.edu

SAMPLE SIZE CONSIDERATIONS FOR THE ANALYSIS OF TIME-VARYING CAUSAL EFFECTS IN STRATIFIED MICRO-RANDOMIZED TRIALS

Walter Dempsey*, *Harvard University*

Peng Liao, *University of Michigan*

Santosh Kumar, *University of Memphis*

Susan A. Murphy, *Harvard University*

Technological advancements have helped overcome obstacles in the delivery of care, making possible delivery of behavioral treatments anytime and anywhere. Increasingly treatment delivery is triggered by predictions of risk or engagement with treatment often designed to impact individuals over a span of time during which subsequent treatments may be provided. We develop an experimental design, the "stratified micro-randomized trial," in which individuals are randomized at times determined by outcomes of past treatment and with randomization probabilities depending on these outcomes. We define both conditional and marginal proximal treatment effects. These effects may be defined over a period of time during which subsequent treatments may be provided. We develop a primary analysis method and associated sample size formulae for testing these effects. This work is motivated by a mobile health smoking cessation study in which randomization probabilities depend on a binary time-varying stress classification and the effect of interest accrues over a period that may include subsequent treatment.

✉ dempsey.walter@gmail.com

BAYESIAN PHASE I/II BIOMARKER-BASED DOSE FINDING FOR PRECISION MEDICINE WITH MOLECULARLY TARGETED AGENTS

Ying Yuan*, *University of Texas MD Anderson Cancer Center*

Beibei Guo, *Louisiana State University*

The optimal dose for treating patients with a molecularly targeted agent may differ according to the patient's individual characteristics, such as biomarker status. In this article, we propose a Bayesian phase I/II dose-finding design to find the optimal dose that is personalized for each patient according to his/her biomarker status. To overcome the curse of dimensionality caused by the relatively large number of biomarkers and their interactions with the dose, we employ canonical partial least squares (CPLS) to extract a small number of components from the covariate matrix containing the dose, biomarkers, and dose-by-biomarker interactions. Using these components as the covariates, we model the ordinal toxicity and efficacy using the latent-variable approach. We quantify the desirability of the dose using a utility function and propose a two-stage dose-finding algorithm to find the personalized optimal dose according to each patient's individual biomarker profile. Simulation studies show that our proposed design has good operating characteristics, with a high probability of identifying the personalized optimal dose.

✉ yyuan@mdanderson.org

ROBUST TREATMENT COMPARISON BASED ON UTILITIES OF SEMI-COMPETING RISKS IN NON-SMALL-CELL LUNG CANCER

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

Thomas A. Murray, *University of Minnesota*

Ying Yuan, *University of Texas MD Anderson Cancer Center*

A design is presented for a randomized clinical trial comparing two second-line treatments, chemotherapy with or without re-irradiation, for recurrent non-small-cell lung cancer. The central question is whether the possible benefit of chemotherapy plus re-irradiation over chemotherapy alone justifies its potential for increasing the risk of toxicity. The co-primary outcomes are the time to disease progression or death, and time to severe toxicity, which are semi-competing risks. A robust conditionally conjugate Bayesian model using piecewise exponential distributions is formulated. A numerical utility function is elicited from the physicians to characterize desirabilities of all possible outcome realizations. A comparative test based on posterior mean utilities is proposed, and a simulation study is presented to evaluate overall test size and power, and sensitivity to the elicited utility function. Guidelines for constructing a design are provided.

✉ peterthall6775@gmail.com

44. MACHINE LEARNING METHODS FOR IMAGING DATA ANALYSIS

IMPROVING PREDICTION ACCURACY THROUGH TRAINING SAMPLE ENRICHMENT FOR HEAVILY UNBALANCED DATA

Peng Huang*, *Johns Hopkins University*

Lung cancer incidence rate is less than 1%/year even among heavy smokers. Extensive publications have shown that computer aided diagnosis (CAD) texture analysis could give high diagnostic accuracy. To apply CAD, a training sample is often selected from historical images to develop diagnostic algorithms. Intuitively, a random sample from the historical data is preferred. However, due to low cancer prevalence, the resulting prediction algorithm could suffer from low positive predictive value (PPV). Although PPV could be increased with increased training sample size, this will substantially increase the computation time. The question is how to select training set with a fixed sample size N that could increase the PPV without losing much in sensitivity. We propose an enriched sampling method to achieve this goal. We will compare our approach with methods using either random training sample or case-control training sample with fixed sample size N.

✉ phuang12@jhmi.edu

COMPUTATIONAL DISCOVERY OF TISSUE MORPHOLOGY BIOMARKER FOR PANCREATIC DUCTAL ADENOCARCINOMA

Pei-Hsun Wu*, *Johns Hopkins University*

Laura D. Wood, *Johns Hopkins University School of Medicine*

Jacob Sarnecki, *Johns Hopkins University*

Ralph H. Hruban, *Johns Hopkins University School of Medicine*

Anirban Maitra, *University of Texas MD Anderson Cancer Center*

Denis Wirtz, *Johns Hopkins University*

Pancreatic ductal adenocarcinoma (PDAC) is one of the deadliest forms of cancer, with an average 5-year survival rate of only 8%. Digital pathology in combination with machine learning provides the opportunities to computational search the tissue morphology patterns in associating with disease outcome. In this work, we developed the computational framework to analyze whole-slide images (WSI) of PDAC patient tissues and identified the prognostic tissue morphology signatures. Based on information from both tissue morphology as well as tissue heterogeneity in tumor and its adjacent area we established a machine learning model with an AUC of 0.94. In sum, our study demonstrates

a pathway to accelerate the discovery of undetermined tissue morphology in association with pathogenesis states and patient outcome for prognosis and diagnosis by utilizing computational approach with digital pathology.

✉ pwu27@jhu.edu

RESIDUAL-BASED ALTERNATIVE PARTIAL LEAST SQUARES FOR FUNCTIONAL LINEAR MODELS

Yue Wang*, *University of North Carolina, Chapel Hill*

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

Hongtu Zhu, University of Texas MD Anderson Cancer Center

The aim of this paper is to develop a residual-based alternative partial least squares estimation (RAPLS) framework of functional partial least squares (fPLS) problems for a large class of functional linear models with functional and scalar predictors. Our RAPLS algorithms integrate iteratively reweighted least squares with an alternative partial least squares (APLS) algorithm to explicitly;handle continuous, categorical, and survival outcomes. We use simulations to illustrate the superior performance of the fPLS estimators over two competing methods, including functional principal component analysis and penalized regression. We apply the proposed RAPLS to the Alzheimer's Disease Neuroimaging Initiative dataset in order to build predictive models for using three dimensional PET data to predict cognitive score and time of conversion from mild cognitive impairment to Alzheimer.

✉ taryue@live.unc.edu

INDIVIDUALIZED MULTILAYER TENSOR LEARNING WITH AN APPLICATION IN IMAGING ANALYSIS

Xiwei Tang*, *University of Virginia*

Xuan Bi, *Yale University*

Annie Qu, *University of Illinois, Urbana-Champaign*

This work is motivated by breast cancer imaging data produced by a multimodality multiphoton optical imaging technique. One unique aspect of breast cancer imaging is that different individuals might have breast imaging at different locations which also creates a technical difficulty. We develop an innovative multilayer tensor learning method to predict disease status effectively through utilizing subject-wise imaging information. In particular, we construct an individualized multilayer model which leverages an additional layer of individual structure of imaging in addition to employing a high-order tensor decomposition shared by populations, which enables us to integrate multimodality imaging data for different profiling of tissue. One major advantage of our approach is that we are able to capture the spatial information of microvesicles observed in different modalities within the same subject. This has medical and clinical significance since identification of microvesicles provides an effective diagnostic tool for early-stage cancer detection.

✉ xtang14@illinois.edu

45. MAKING SENSE OF WHOLE GENOME SEQUENCING DATA IN POPULATION SCIENCE: STATISTICAL CHALLENGES AND SOLUTIONS

ANALYSIS OF WHOLE GENOME SEQUENCING ASSOCIATION STUDIES: CHALLENGES AND OPPORTUNITIES

Xihong Lin*, *Harvard University*

Whole genome sequencing data and different types of genomics data have become rapidly available. Two large ongoing whole genome sequencing programs (Genome Sequencing Program (GSP) of NHGRI and Trans-omics for Precision Medicine Program (TOPMed) of NHLBI) plan to sequence 300,000-350,000 whole genomes. These massive genetic and genomic data present many exciting opportunities as

well as challenges in data analysis and result interpretation. In this talk, I will discuss several statistical and computational methods for analysis of whole-genome sequencing association studies.

✉ xlin@hsph.harvard.edu

WE DID NOT SEE THIS IN GWAS: UNDERSTANDING AND FIXING UNFAMILIAR PROBLEMS IN ASSOCIATION ANALYSES, WHEN POOLING WHOLE GENOME SEQUENCE DATA FROM MULTIPLE STUDIES

Kenneth M. Rice*, *University of Washington*

Xiuwen Zheng, *University of Washington*

Stephanie Gogarten, *University of Washington*

Tamar Sofer, *University of Washington*

Cecelia Laurie, *University of Washington*

Cathy Laurie, *University of Washington*

Bruce Weir, *University of Washington*

Tim Thornton, *University of Washington*

Adam Szpiro, *University of Washington*

Jen Brody, *University of Washington*

Large-scale association analyses are now underway, using whole genome sequence (WGS) data on thousands of participants. Unlike earlier GWAS, where data were combined by meta-analysis of summary statistics, participant-level WGS data from multiple studies is typically pooled into a single analysis. While there are good reasons for this approach to WGS, we describe how it can lead to false-positive results when the equivalent GWAS-style approach would not. Specifically, we consider the impact of differential phenotype variances by study (due to e.g. different protocols) and its interplay with adjustment for relatedness across studies (e.g. allowing random variability proportional to a kinship or other genetic relatedness matrix). As well as explaining why these issues lead to difficulties in WGS where they did not for GWAS, we describe methods suitable for WGS work - available in straightforward and freely-available software - that used pooled data and provide appropriate control of false-positive results. For both single-SNP and region-based analyses, the problems and their solutions are illustrated with several examples from the NHLBI's TOPMed Program.

✉ kenrice@u.washington.edu

STATISTICAL METHODS AND TOOLS FOR WHOLE-GENOME SEQUENCING DATA ANALYSIS OF 100,000 SAMPLES

Seunggeun Lee*, *University of Michigan*

The rapid decrease in sequencing cost has enabled to sequence large numbers of whole genomes to find the genetic basis of complex diseases. TOPMed is currently attempting to sequence 100,000 whole genomes with heart lung and blood phenotypes. This large dataset provides a great opportunity for new discoveries; at the same time poses statistical and computational challenges. In this talk, I will first introduce cloud-based tools developed in TOPMed informatics research center (IRC). These tools provide an easy-to-use interface and scalable implementation of the current best practice in association tests. In the second part of the talk, I will introduce a new statistical method, SAIGE, that can analyze 100,000 samples for binary phenotypes with adjusting for family relatedness and case-control imbalance. SAIGE uses the saddlepoint approximation to adjust for case-control imbalance at the top of the Generalized Mixed Model method. In addition, it uses state of art optimization techniques to analyze $>$ 100,000 samples. These tools and methods will enable large-scale analysis and be useful not only in TOPMed but also other large biobank-based studies.

✉ leeshawn@umich.edu

A SEMI-SUPERVISED APPROACH FOR PREDICTING TISSUE SPECIFIC FUNCTIONAL EFFECTS OF NONCODING VARIATION

Iuliana Ionita-Laza*, *Columbia University*

Zihuai He, *Columbia University*

Understanding the functional consequences of genetic variants in noncoding regions is difficult. Projects like ENCODE and Roadmap provide various epigenetic features genome-wide in over a hundred tissues/cell types. Over the past few years, several unsupervised approaches have been proposed to integrate these epigenetic features to predict the functional effects of variants. While the unsupervised approaches can be advantageous when the amount and quality of labeled data are limited, supervised methods are expected to perform better. As more experimental data (massively parallel reporter assay, CRISPR/Cas9) across different tissues/cell types become available, it becomes feasible to jointly utilize both experimentally confirmed regulatory variants and large number of functional annotations to predict functional effects of genetic variants. We propose here a semi-supervised approach, GenoNet, to jointly utilize experimentally confirmed regulatory variants, millions of unlabeled variants genome-wide, and more than a thousand cell/tissue specific functional annotations on each variant to predict functional effects of noncoding variants.

✉ ii2135@columbia.edu

46. MACHINE LEARNING METHODS FOR PRECISION MEDICINE

RECENT DEVELOPMENTS IN REINFORCEMENT LEARNING FOR DECISION SCIENCE

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

In this talk, we develop novel reinforcement learning methodology for discovery of tailored action regimes for complex decision making in a variety of settings. We also consider challenges arising from having multiple, conflicting outcomes to optimize.

✉ kosorok@bios.unc.edu

TARGETED MACHINE LEARNING FOR PRECISION MEDICINE

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

Alex Luedtke, *Fred Hutchinson Cancer Research Center*

Targeted minimum loss estimation (TMLE), which provides a general template for the construction of asymptotically efficient plug-in estimators of a target estimand for infinite dimensional models. TMLE involves maximizing a parametric likelihood along a so-called least favorable parametric model through an initial estimator (e.g., ensemble super-learner (SL)) of the relevant functional of the data distribution. The asymptotic normality and efficiency of the TMLE relies on the asymptotic negligibility of a second-order term. In observational studies this requires the initial estimator to converge at a rate faster than n-1/4. We propose a new estimator, the Highly Adaptive LASSO (HAL), of the data distribution and its functionals that converges at a sufficient rate regardless of the dimensionality of the data/model, under almost no additional regularity. This allows us to propose a general TMLE that is asymptotically efficient in great generality. Wewill demonstrate a super learner of the optimal dynamic treatment regime itself and a TMLE of the mean counterfactual outcome under the optimal dynamic regime. under resource constraints.

✉ laan@berkeley.edu

GENERALIZED RANDOM FORESTS

Stefan Wager*, *Stanford University*

Julie Tibshirani, *Palantir Technologies*

Susan Athey, *Stanford University*

We propose generalized random forests, a method for non-parametric statistical estimation based on random forests (Breiman, 2001) that can be used to fit any quantity

of interest identified as the solution to a set of local moment equations. Our method operates at a particular point in covariate space by considering a weighted set of nearby training examples; however, instead of using classical kernel weighting functions that are prone to a strong curse of dimensionality, we use an adaptive weighting function derived from a forest designed to express heterogeneity in the specified quantity of interest. We propose a flexible, computationally efficient algorithm for growing generalized random forests, develop a large sample theory for our method showing that our estimates are consistent and asymptotically Gaussian, and provide an estimator for their asymptotic variance that enables valid confidence intervals. We use our approach to develop new methods for three statistical tasks: non-parametric quantile regression, conditional average partial effect estimation, and heterogeneous treatment effect estimation via instrumental variables.

✉ swager@stanford.edu

OPTIMAL INDIVIDUALIZED TREATMENTS WHEN MEASURING COVARIATES IS EXPENSIVE

Alex Luedtke*, *Fred Hutchinson Cancer Research Center*

Consider an individualized treatment setting in which the covariates that best inform decision making are expensive to measure and resources are limited so that these covariates cannot be measured on everyone. Therefore, one wishes to optimize both the covariates that should be measured on each subject and also the treatment decision that should be made given the measured covariates. One can frame this problem as a resource-constrained sequential decision problem, where each decision point corresponds to a decision regarding either which covariate to measure or which treatment to administer. We derive finite-sample guarantees for a decision rule estimation strategy in this setting, and develop confidence intervals for the mean outcome that would be observed if the optimal decision strategy were implemented in the population. These confidence intervals are valid in non-regular settings where the optimal decision strategy is non-unique.

✉ aluedtke@fredhutch.org

47. MULTI-OMICS AND GRAPHICAL MODELS FOR PRECISION MEDICINE

HETEROGENEOUS RECIPROCAL GRAPHICAL MODELS

Yuan Ji*, *NorthShore University HealthSystem and University of Chicago*

We develop novel hierarchical reciprocal graphical models to infer gene networks from heterogeneous data. In the case of data that can be naturally divided into known groups, we propose to connect graphs by introducing a hierarchical prior across group-specific graphs, including a correlation on edge strengths across graphs. Thresholding priors are applied to induce sparsity of the estimated networks. In the case of unknown groups, we cluster subjects into subpopulations and jointly estimate cluster-specific gene networks, again using similar hierarchical priors across clusters. We illustrate the proposed approach by simulation studies and three applications with multiplatform genomic data for multiple cancers.

✉ koaeraser@gmail.com

CONSTRUCTING TUMOR-SPECIFIC GENE REGULATORY NETWORKS BASED ON SAMPLES WITH TUMOR PURITY HETEROGENEITY

Pei Wang*, *Icahn School of Medicine at Mount Sinai*

Francesca Petralia, *Icahn School of Medicine at Mount Sinai*

Li Wang, *Icahn School of Medicine at Mount Sinai*

Jie Peng, *University of California, Davis*

Tumor tissue samples often contain an unknown fraction of normal cells. This problem well known as tumor purity heterogeneity (TPH) was recently recognized as a severe issue in omics studies. Specifically, if TPH is ignored when

inferring co-expression networks, edges are likely to be estimated among genes with mean shift between normal and tumor cells rather than among gene pairs interacting with each other in tumor cells. To address this issue, we propose TSNet --- a new method which constructs tumor-cell specific gene/protein co-expression networks based on gene/protein expression profiles of tumor tissues. The advantage of TSNet over existing methods ignoring TPH is illustrated through extensive simulation examples. We then apply TSNet to estimate tumor specific co-expression networks based on breast cancer expression profiles. We identify novel co-expression modules and hub structure specific to tumor cells.

✉ pei.wang@mssm.edu

JOINT SKELETON ESTIMATION OF MULTIPLE DIRECTED ACYCLIC GRAPHS FOR HETEROGENEOUS POPULATION

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

Wei Sun, *Fred Hutchinson Cancer Research Center*

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

The directed acyclic graph (DAG) is a powerful tool to model the interactions of high-dimensional variables. While estimating edge directions in a DAG often requires interventional data, one can estimate the skeleton of a DAG using observational data. In real data analyses, the samples of the high-dimensional variables may be collected from a mixture of multiple populations. Each population has its own DAG while the DAGs across populations may have significant overlap. In this paper, we propose a two-step approach to jointly estimate the DAG skeletons of multiple populations while the population origin of each sample may or may not be labeled. In particular, our method allows a probabilistic soft label for each sample, which can be easily computed and often leads to more accurate skeleton estimation than hard labels. Compared with separate estimation of skeletons for each population, our method is more accurate and robust to labeling errors. Simulation studies are performed to demonstrate the performance of the new method. Finally, we apply our method to analyze gene expression data from breast cancer patients of multiple cancer subtypes.

✉ liuoo@unc.edu

BAYESIAN MULTI-LAYERED GAUSSIAN GRAPHICAL MODELS

Min Jin Ha*, *University of Texas MD Anderson Cancer Center*

Francesco Stingo, *University of Florence*

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

Simultaneous modeling of data arising from multiple ordered layers provides insight into the holistic picture of the interactive system and the flow of information. Chain graphs have been used to model the layered architecture of networks where the vertices can be naturally partitioned into ordered layers that exhibit undirected and directed acyclic relations within and between the layers. We develop a multi-layered Gaussian graphical model (mlGGM) to investigate conditional independence structures in probabilistic chain graphs. Our proposed model uses a Bayesian node-wise selection framework that coherently accounts for dependencies in the mlGGM. Using Bayesian variable selection strategies for each of the node-wise regressions allows for flexible modeling, sparsity and incorporation of edge-specific prior knowledge. Through simulated data generated from various scenarios, we demonstrate that our node-wise regression method outperforms other related multivariate regression-based methodologies. We apply mlGGM to identify integrative networks for key signaling pathways in kidney cancer and dynamic signaling networks using longitudinal proteomics data in breast cancer.

✉ mjha@mdanderson.org

48. RECENT ADVANCES IN PROPENSITY SCORE ANALYSIS

UTILIZING PROPENSITY SCORE METHODOLOGY IN THE OBSERVATIONAL STUDIES WITH "BIG DATA" IN THE REGULATORY SETTINGS

Lilly Q. Yue, U.S. Food and Drug Administration*

Formulated by Rosenbaum and Rubin, the propensity score methodology is a versatile statistical technique used mainly in observational (non-randomized) studies for improving treatment comparison by adjusting for up to a relatively large number of potentially confounding covariates in estimating the treatment effect on outcomes. Since its formulation, the methodology has been widely used in areas such as epidemiological and social science studies. There has been an increased interest in applying this methodology to non-randomized clinical studies, including those designed to support regulatory applications for marketing medical products. Recently, the methodology has been utilized in the regulatory clinical studies leveraging "big data". This presentation discusses important things to consider in the application of methodology in the regulatory settings, and highlights a fundamental distinction between an exploratory study of general research and a regulatory confirmatory study in applying the methodology.

✉ lilly.yue@fda.hhs.gov

SUBGROUP BALANCING PROPENSITY SCORE

Fan Li*, *Duke University*

Jing Dong, *Industrial and Commercial Bank of China*

Junni Zhang, *Peking University*

We investigate the estimation of subgroup treatment effects with observational data. Existing propensity score matching and weighting methods are mostly developed for estimating overall treatment effect. Although the true propensity score should balance covariates for the subgroup populations, the estimated propensity score may not balance covariates for the subgroup samples. We propose the subgroup balancing propensity score (SBPS) method, which selects, for each subgroup, to use either the overall sample or the subgroup sample to estimate propensity scores for units within that subgroup, in order to optimize a criterion accounting for a set of covariate-balancing conditions for both the overall sample and the subgroup samples. We develop a stochastic search algorithm for the estimation of SBPS when the number of subgroups is large. We demonstrate through simulations that the SBPS can improve the performance of propensity score matching in estimating subgroup treatment effects. We then apply SBPS to the Italy Survey of Household Income and Wealth data to estimate the treatment effects of having debit card on household consumption for different income groups.

✉ fli@stat.duke.edu

PROPENSITY SCORE ANALYSIS WITH COMPLEX SURVEY DATA: WHEN TREATMENT ASSIGNMENT AND TREATMENT EFFECTS VARY ACROSS STRATA AND CLUSTERS

Trang Q. Nguyen*, *Johns Hopkins Bloomberg School of Public Health*

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

Nationally representative samples are often used to estimate effects of exposures on health/education outcomes, often using propensity score (PS) methods to balance confounders. How PS methods apply to complex samples is an open research area. We address how strata/clusters should be handled, assuming treatment assignment and treatment effects vary across strata/clusters. In a setting with several strata and many clusters, we let strata differ systematically and clusters differ randomly in covariate distribution, covariates' influence on treatment assignment, treatment effect, and covariates' effect modification. Using PS weighting to estimate the population average treatment effect, we

investigate 3 strategies: ignoring strata (naive); treating stratum indicators as covariates; covariate balancing within strata and letting outcome models vary by stratum (stratified analysis). We found that random cluster variations do not induce bias. With systematic treatment effect variation across strata, the naive method is biased. When covariates' influence on treatment assignment varies across strata, the strata-as-covariates method is also biased, while stratified analysis remains unbiased.

✉ tnguye28@jhu.edu

49. CAUSAL INFERENCE AND EPIDEMIOLOGICAL METHODS

A NOVEL NON-PARAMETRIC METHOD FOR ORDINAL PROPENSITY SCORE MATCHING AND STRATIFICATION

Thomas James Greene*, *University of Texas Health Science Center at Houston*

Stacia M. DeSantis, *University of Texas Health Science Center at Houston*

Michael D. Swartz, *University of Texas Health Science Center at Houston*

Currently, all methods of conducting ordinal treatment propensity score matching rely on matching using the linear predictor from the proportional odds (PO) model. In practice the PO assumption may not hold, and propensity score matching techniques for ordinal treatment can be biased. This issue motivates the development of a new flexible method of propensity matching and stratification which does not require strong parametric assumptions such as proportional odds. The proposed method uses non-linear least squares to fit a one-parameter power function to the cumulative distribution function (CDF) of the generalized propensity score (GPS) vector and is known as the GPS-CDF method. This estimated parameter acts as a balancing score that can be used to assess subject similarity. Similar subjects who received different levels of treatment are then matched or stratified based on their value. Matching or stratifying on a balancing score enables proper estimation the average treatment effect. Simulation results show desirable operating characteristics, including minimal bias, adequate coverage probability, and removal of covariate imbalance.

✉ jay.greeneiv@gmail.com

OPTIMAL TRADEOFFS IN MATCHED DESIGNS FOR OBSERVATIONAL STUDIES

Samuel D. Pimentel*, *University of California, Berkeley*

Rachel R. Kelz, *University of Pennsylvania*

An effective matched design for causal inference in observational data must achieve several goals, including balancing covariate distributions marginally, ensuring units within individual pairs have similar values on key covariates, and using a sufficiently large sample from the raw data. Yet optimizing one of these goals may force a less desirable result on another. We address such tradeoffs from a multi-objective optimization perspective by creating matched designs that are Pareto optimal with respect to two goals. We provide tools for generating representative subsets of Pareto optimal solution sets and articulate how they can be used to improve decision-making in observational study design. We illustrate the method in reanalysis of a large surgical outcomes study comparing outcomes of patients treated by US-trained surgeons and of patients treated by internationally-trained surgeons. Formulating a multi-objective version of the problem helps us evaluate the cost of balancing an important variable in terms of two other design goals, average closeness of matched pairs on a multivariate distance and size of the final matched sample.

✉ spi@berkeley.edu

EVALUATING THE PERFORMANCE OF BALANCING SCORES USING THE ANCOVA APPROACH FOR ESTIMATING AVERAGE TREATMENT EFFECTS IN OBSERVATIONAL STUDIES: A SIMULATION STUDY

Woon Yuen Koh*, *University of New England*

Chunhao Tu, *University of New England*

We conducted a simulation study to evaluate the performance of five balancing scores using the Analysis of Covariance (ANCOVA) approach, for adjusting bias in estimating average treatment effects (ATE) in observational studies. The five balancing scores which we used as the covariate(s) in the ANCOVA model were (1) propensity score (P), (2) prognostic score (G), (3) propensity score estimated by prognostic score (PG), (4) prognostic score estimated by propensity score (GP), and (5) both propensity and prognostic scores (P&G). The results of the five balancing scores using the ANCOVA approach were compared to the results of the classic regression approach, which included all observed covariates as the predictors (X). Simulation results showed that balancing scores P, GP, and P&G had the smallest bias and MSE when the outcome variable and the observed covariates were linearly associated, and PG had the smallest bias and MSE when the association was nonlinear.

✉ wkoh@une.edu

PRINCIPAL STRATIFICATION FOR LONGITUDINAL DATA IN ENVIRONMENTAL TRIALS

Joshua P. Keller*, *Johns Hopkins University*

Roger D. Peng, *Johns Hopkins University*

Elizabeth C. Matsui, *Johns Hopkins University*

Randomized trials of indoor air quality interventions have shown promise for identifying approaches for improving indoor air pollution and reducing adverse respiratory symptoms in children. A recent two-city trial targeting mouse allergens compared an active intervention of professional pest management services against education alone. An intent-to-treat analysis found no significant difference in repeated measures of respiratory symptoms between treatment and control. However, the intent-to-treat approach ignores available information about the impact of treatment assignment on actual mouse allergen concentrations. Using a principal stratification approach, we estimate the potential health benefit among subjects who would have a reduction in mouse allergen levels when assigned to treatment and among those who would see no notable change in mouse allergen levels under treatment. This analysis demonstrates the applicability of principal stratification for environmental trials with continuous exposures and repeated observations.

✉ jkelle46@jhu.edu

EFFICIENT COMPUTATION OF THE JOINT PROBABILITY OF MULTIPLE GERMLINE MUTATIONS FROM PEDIGREE DATA

Thomas Madsen*, *Harvard School of Public Health*

Danielle Braun, *Dana-Farber Cancer Institute*

Lorenzo Trippa, *Dana-Farber Cancer Institute*

Giovanni Parmigiani, *Dana-Farber Cancer Institute*

The Elston-Stewart peeling algorithm enables estimation of a patient's probability of harboring germline mutations based on pedigree data. The algorithm combines information about a patient's family history of disease, the prevalence and penetrance of a pre-specified set of mutations, and principles of Mendelian inheritance and probability theory. It serves as the computational backbone of many risk prediction tools. However, it remains limited to the analysis of a small number of mutations because its computing time grows exponentially with the number of genetic loci considered. We propose a novel, approximate version of this algorithm which scales polynomially in the number of loci. This enables risk prediction tools to include

many pathogenic germline mutations. The algorithm creates a trade-off between accuracy and speed, and allows the user to control this trade-off. We illustrate our approximation on simulated data from an extended version of BRCAPRO, a risk prediction model for estimating the carrier probabilities of BRCA1 and BRCA2 mutations. Results show that the loss of accuracy is negligible. Exact bounds on the approximation error are also discussed.

✉ tmadsen@g.harvard.edu

A DYNAMIC MODEL FOR EVALUATION OF BIAS OF ESTIMATES OF INFLUENZA VACCINE EFFECTIVENESS FROM OBSERVATIONAL STUDIES

Kylie E. C. Ainslie*, *Emory University*

Michael Haber, *Emory University*

As influenza vaccination is now widely recommended, observational studies based on patients with acute respiratory illness (ARI) seeking medical care remain the only option for estimating influenza vaccine effectiveness (IVE). We developed a dynamic probability model for the evaluation of bias of IVE estimates from four commonly used observational study designs: active surveillance cohort (ASC), passive surveillance cohort, test-negative (TN), and traditional case-control. The model includes two covariates (health status and health awareness), which may affect the probabilities of vaccination, developing ARI, and seeking medical care. We consider two outcomes of interest: symptomatic influenza (SI) and medically-attended influenza (MAI). Our results suggest that when the outcome of interest is SI, ASC studies produce unbiased estimates, except when health status influences the probability of influenza ARI. When vaccination affects the probability of non-influenza ARI, IVE estimates from TN studies may be severely biased, while IVE estimates from cohort studies are unbiased. However, TN estimates are unbiased against MAI when some sources of bias are present.

✉ kylie.ainslie@emory.edu

50. EPIDEMIOLOGICAL METHODS

APPLICATION OF REGRESSION ANALYSIS ON TEXT-MINING DATA ASSOCIATED WITH AUTISM SPECTRUM DISORDER FROM TWITTER: A PILOT STUDY

Chen Mo*, *Georgia Southern University*

Jingjing Yin, *Georgia Southern University*

Isaac Chun-Hai Fung, *Georgia Southern University*

Zion Tse, *University of Georgia*

Social media has become a popular resource of health data analysis. Mathematics and computation techniques are challenging to public health practitioners when using the massive data from social media. Besides, it is difficult to interpret results from traditional machine learning techniques. This study proposes a simple new solution by regressing the primary outcome of interest (e.g., number of retweets of a tweet or whether a tweet contains certain keywords) on the frequency of common terms appeared in the tweet. This method reduces the term matrix based on the fitted regression scores, such as relative risk or odds ratio. It also solves the data sparsity issue and transforms text data into continuous summary scores. It would be easier to perform data analysis on social media data and interpret the results using the proposed scores. We used a twitter data of Autism Spectrum Disorder (ASD) and applied regression models for analysis, including poisson model, hurdle model and logistic model with model selection based on the Youden index. We found that the terms with significant results are generally present the key factors associated with ASD in the existing literature.

✉ cm06957@georgiasouthern.edu

A PSEUDOLIKELIHOOD METHOD FOR ESTIMATING MISCLASSIFICATION PROBABILITIES WHEN TRUE OUTCOMES ARE PARTIALLY OBSERVED

Philani Brian Mpofu*, *Indiana University Purdue University, Indianapolis*

Giorgos Bakoyannis, *Indiana University Purdue University, Indianapolis*

Constantin Yiannoutsos, *Indiana University Purdue University, Indianapolis*

The problem of misclassification of binary outcome data has been extensively studied, with remedies such as internal validation sampling using a gold standard diagnostic procedure having been proposed in order to correctly perform statistical inference. However, due to financial and other constraints, internal validation sampling is not always feasible. It is, thus, very important to estimate misclassification probabilities from studies with internal validation and use these estimates to correct for misclassification in studies without internal validation. For the first task, we propose a computationally efficient pseudo-likelihood method for estimating misclassification probabilities when true outcomes are partially observed. We then show how these estimates can be applied in an external study to correct for misclassification. We show the root-n consistency of the estimator and its asymptotic normality. We also derive a closed-form variance estimator that accounts for all sources of uncertainty. We illustrate the method using data from a HIV cohort study in sub-Saharan Africa to estimate death under-reporting in settings where internal validation has not been performed.

✉ phmpofu@umail.iu.edu

CORRECTING FOR RISK FACTOR MISCLASSIFICATION IN THE PARTIAL POPULATION ATTRIBUTABLE RISK

Benedict Wong*, *Harvard T.H. Chan School of Public Health*

Donna Spiegelman, *Harvard T.H. Chan School of Public Health*

Molin Wang, *Harvard T.H. Chan School of Public Health*

Estimation of the partial population attributable risk (pPAR) has become an important goal in epidemiologic research, because it describes the proportion of disease cases that could be prevented if a set of exposures were entirely eliminated from a target population, when the distributions of other risk factors, possibly unmodifiable, exist but do not change as a result of some intervention. In epidemiological studies, categorical covariates are often misclassified, and this results in biased estimates of the risk factor prevalences and the relative risks. We present methods for obtaining corrected point and interval estimates of the pPAR after adjusting for misclassification, and we apply these methods to data from the Health Professionals Follow-Up Study.

✉ wong01@fas.harvard.edu

REGRESSION ANALYSIS OF TEMPORAL BIOMARKER EFFECTS UNDER NESTED CASE-CONTROL STUDIES

Yiding Zhang*, *University of Massachusetts, Amherst*

Jing Qian, *University of Massachusetts, Amherst*

Susan E. Hankinson, *University of Massachusetts, Amherst*

Nested case-control (NCC) study design, which is cost-effective, has been widely used in large prospective epidemiologic studies. Biomarkers exhibiting time-varying associations with disease outcome occur frequently in clinical studies. To analyze data from NCC study design with temporal biomarker effects, we adopt a generalized Cox model which allows time-varying covariate effects. We develop a computationally efficient estimating procedure by integrating martingale-based estimating equations with inverse probability weighting. The proposed estimating procedure is able to correct the sampling bias in NCC study design and thus yield unbiased estimates for the time-varying covariate effects. We also develop a resampling

procedure for variance estimation, by restoring the complex correlation structure induced by sampling in NCC. We establish the asymptotic properties of the resulting estimators, including uniform consistency and weak convergence. Simulation studies demonstrate nice finite sample performance of the proposed procedures. The proposed methods are applied to the Nurses' Health Study for evaluating association of hormone biomarkers with breast cancer risk.

✉ yidingzhang@schoolph.umass.edu

A NOVEL GOODNESS-OF-FIT BASED TWO-PHASE SAMPLING DESIGN FOR STUDYING BINARY OUTCOMES

Le Wang*, *University of Pennsylvania*

Xinglei Chai, *University of Pennsylvania*

Yong Chen, *University of Pennsylvania*

Jinbo Chen, *University of Pennsylvania*

In biomedical cohort studies for assessing the association between an outcome and a set of covariates, it is common that some covariates can only be measured on a subgroup of study subjects. An important design question is which subjects to select into the subgroup towards increased statistical efficiency for association analyses. When the outcome is binary, one may adopt a case-control or a balanced design where cases and controls are further matched on a small number of complete discrete covariates. While the latter achieves success in estimating odds ratio (OR) for the matching covariates, to our best knowledge, similar two-phase designs are not available to increase statistical efficiency for other covariates, especially the incomplete ones. To this end, we propose a novel sampling scheme that oversamples cases and controls with worse goodness-of-fit based on an external model relating outcome and complete covariates and further matches them on complete covariates similarly to the balanced design. We developed a pseudo-likelihood method for OR estimation and found that it led to reduced asymptotic variances of OR estimates through simulations and a real cohort study.

✉ lwang0217@gmail.com

ON OPTIMAL TWO-PHASE DESIGNS

Ran Tao*, *Vanderbilt University Medical Center*

Donglin Zeng, *University of North Carolina, Chapel Hill*

Danyu 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. Previous research on two-phase studies has largely focused on the inference procedures rather than the design aspects. We investigate the design efficiency of two-phase studies, defined as the semiparametric efficiency bound of estimating the regression coefficients of expensive covariates. We calculate the design efficiency of general two-phase studies, where the outcome variable can be continuous, discrete, or time-to-event, and the second-phase sampling can depend on the first-phase data in any manner. We develop optimal two-phase designs, which can be substantially more efficient than existing designs. We evaluate the efficiencies of the optimal designs and existing ones through extensive simulation studies. We provide an application to the National Heart, Lung, and Blood Institute Exome Sequencing Project.

✉ r.tao@vanderbilt.edu

5I. INFECTIOUS DISEASE MODELS

A BAYESIAN GENERALIZED ADDITIVE MODEL FOR GROUP TESTING DATA

Christopher S. McMahan*, *Clemson University*

Yan Liu, *University of Nevada, Reno*

Joshua M. Tebbs, *University of South Carolina*

Christopher R. Bilder, *University of Nebraska-Lincoln*

Colin M. Gallagher, *Clemson University*

For screening for infectious diseases, group testing has proven to be a cost-effective alternative to one-at-a-time testing, with cost savings being realized through assaying

pooled biospecimen; urine, blood, etc. Within this venue, a common goal is to conduct disease surveillance, which often involves relating the individuals' true disease statuses to covariate information through a binary regression model. To this end, several authors have developed binary regression methodologies specifically designed to accommodate the complex structure of the data observed from group testing strategies, to include nonparametric techniques. Extending and generalizing all this previous work, we propose a Bayesian generalized additive model. This model can be used to analyze data arising from any group testing procedure with the goal of estimating multiple unknown smooth functions of the covariates, standard linear effects in other predictor variables, and even multiple assay accuracy probabilities. The finite sample performance of the proposed methodology is extensively evaluated through Monte Carlo simulation studies, and it is used to analyze chlamydia data collected in Iowa.

✉ mcmaha2@clemson.edu

A GENERAL MULTIVARIATE BAYESIAN REGRESSION MODEL FOR GROUP TESTING DATA

Paul J. Cubre*, *Clemson University*

Christopher S. McMahan, *Clemson University*

Yingbo Li, *Clemson University*

In clinical laboratories throughout the United States and elsewhere, multiplex assays are being adopted for the purpose of screening for infectious diseases. These diagnostic tests, unlike their predecessors, test for multiple infectious agents simultaneously. In general, these assays result in reducing both the time and cost of testing. In order to further reduce cost in high volume settings, many diagnostic laboratories are adopting group testing in conjunction with multiplex assays. Group testing reduces cost by assaying pooled specimen rather than testing the individual specimen one-by-one. The data observed from multiplex group testing is extremely complex and to date regression methodologies within this venue have been underdeveloped. In this work we develop a general Bayesian regression methodology that can be used to analyze data arising from any multiplex group testing protocol. This approach jointly models the latent polychromatic response and can be used to estimate the assay accuracy probabilities. The performance of the proposed approach is demonstrated through simulation and is illustrated using chlamydia data collected by the Iowa public health department.

✉ pcubre@clemson.edu

BAYESIAN REGRESSION ANALYSIS OF MULTIPLE-INFECTION GROUP TESTING DATA WITH A CONSIDERATION OF DILUTION EFFECTS

Juexin Lin*, *University of South Carolina*

Dewei Wang, *University of South Carolina*

Group testing has been widely used as a cost-effective procedure in large-scale screening for an infectious disease. The recent development of multiplex assays has extended the content from a single infection to multiple infections which yields group testing data of more complex structures. Existing statistical analysis of such data either do not consider individual covariate information or cannot incorporate possible retests on suspicious individuals. In this article, we build a comprehensive Bayesian regression framework that can achieve both. Our framework uses a copula to jointly model all the infections while being able to produce interpretable marginal inference for each infection separately. In addition, our framework is able to estimate the assay sensitivity and specificity for each infection and to detect possible dilution effects that caused by pooling. We illustrate our methodology through simulation and a chlamydia and gonorrhea data collected from the Infertility Prevention Project.

✉ juexin@email.sc.edu

STUDYING THE PATTERN OF TEMPORAL ASSOCIATIONS BETWEEN RARE DISEASE INCIDENCE AND METEOROLOGICAL FACTORS USING A BAYESIAN CONDITIONAL POISSON MODEL WITH A GAUSSIAN PROCESS PRIOR OVER THE DISTRIBUTED LAG COEFFICIENTS

James L. Crooks*, *National Jewish Health*

Conditional Poisson models (Armstrong et al., 2014), like conditional logistic models, were developed to allow automatic control of time-invariant, slowly varying, or cyclic confounders in count data time series. Whereas conditional logistic models can accommodate individual-level confounders, conditional Poisson models are appropriate for situations where confounders are known at a group level. We apply conditional Poisson models to laboratory-confirmed cases of Tularemia in domestic and wild animals in the United States over the years 2008-2016. Incidence of Tularemia in animals is the main source of infection in humans and has increased in parallel with human cases over the past decade, possibly influenced by climatic factors. To understand these climatic influences, we study associations between state-level daily case counts and meteorological factors up to 360 days prior to the case report. Specifically, we model lagged effects in ten-day increments using a distributed lag structure with a Gaussian process prior over the lagged coefficients to induce temporal smoothing. We implement the model in Stan.

✉ CrooksJ@NJHealth.org

PAIRWISE ACCELERATED FAILURE TIME MODELS FOR INFECTIOUS DISEASE TRANSMISSION WITHIN AND BETWEEN HOUSEHOLDS

Yushuf Sharker*, *Yale School of Public Health*

Eben Kenah, *The Ohio State University*

Kenah(2011) showed that parametric survival analysis can be used to handle dependent happenings in infectious disease transmission data by taking ordered pairs of susceptible-infected individuals as the units of analysis. In this approach, the failure time the contact interval, the time from the onset of infectiousness in an individual i to infectious contact from i to individual j, where an infectious contact is sufficient to infect j if he/she is susceptible. These methods assumed the same contact interval distribution in all pairs. We generalize pairwise survival analysis in two ways: First, introduce a pairwise accelerated failure time model in which the rate parameter of the contact interval distribution depends on covariates associated with infectiousness in i and susceptibility in j. Second, we show how internal infections (within a household) and external infections (sourced from outside) can be handled simultaneously. In simulations, we show that these methods produce valid point and interval estimates of transmission probabilities and rate ratios. We use these methods to analyze influenza A(H1N1) surveillance data from Los Angeles County during the 2009 pandemic.

✉ mayushuf@gmail.com

52. MULTIVARIATE SURVIVAL ANALYSIS

SPEARMAN'S RANK CORRELATION ADJUSTING FOR COVARIATES IN BIVARIATE SURVIVAL DATA

Svetlana K. Eden*, *Vanderbilt University*

Chun Li, *Case Western Reserve University*

Bryan E. Shepherd, *Vanderbilt University*

Many studies are interested in measuring associations between two right-censored time-to-event variables, sometimes called bivariate survival data. For example, researchers may want to assess associations between the times to cardiovascular disease for patients and their parents, or between times to events in twins. We develop a rank-based method to measure associations with and without adjusting for covariates. Our method fits separate semi-parametric models for the times to events conditional on covariates, obtains probability scale residuals (PSRs; Shepherd, Li, Liu [2016]) from these fitted models, and then computes the correlation of the PSRs. We show that without covariates, the correlation of PSRs equals Spearman's rank

correlation for censored data. With covariates, the method is a natural extension of Spearman's correlation to permit covariate adjustment and censoring. We propose ways to estimate the variance of our estimators and demonstrate their performance using simulations. We illustrate by investigating the association between times from treatment initiation to viral failure and regimen change among HIV-positive persons.

✉ svetlana.eden@vanderbilt.edu

COPULA-BASED SEMIPARAMETRIC SIEVE MODEL FOR BIVARIATE INTERVAL-CENSORED DATA, WITH AN APPLICATION TO STUDY AMD PROGRESSION

Tao Sun*, *University of Pittsburgh*

Wei Chen, *University of Pittsburgh*

Ying Ding, *University of Pittsburgh*

This research is motivated by discovering genetic causes for the progression of a bilateral eye disease, Age-related Macular Degeneration (AMD), where the primary outcomes, progression times to late AMD, are bivariate and interval-censored. We develop a flexible copula-based semiparametric approach for modeling and testing bivariate interval-censored data. Specifically, the joint likelihood is modeled through a two-parameter Archimedean copula, which can flexibly characterize the dependence structure between two margins. The marginal distributions are modeled through a semiparametric transformation model using sieves, with the proportional hazards or odds model being a special case. We propose a two-step maximum likelihood estimation procedure and develop a computationally efficient score test, which is suitable for large-scale testing as we consider here. We establish asymptotic properties of the proposed estimator. Extensive simulations are conducted to evaluate the performance of the proposed method in finite samples. Finally, we apply our method to a genome-wide analysis of AMD progression, to identify susceptible risk variants for the disease progression.

✉ suntaojj@gmail.com

EM ALGORITHMS FOR FITTING MULTISTATE CURE MODELS

Lauren J. Beesley*, *University of Michigan*

Jeremy M. G. Taylor, *University of Michigan*

Multistate cure models are multistate models in which transitions into one or more of the states cannot occur for a fraction of the population. In this talk, we present an Expectation-Maximization (EM) algorithm for fitting the multistate cure model using maximum likelihood. The proposed algorithm makes use of a weighted likelihood representation allowing it to be easily implemented with standard software and can incorporate either parametric or nonparametric baseline hazards for the state transition rates. A common complicating feature in cancer studies is that the follow-up time for recurrence may differ from the follow-up time for death. Additionally, we may have missingness in the covariates. We propose a Monte Carlo EM (MCEM) algorithm for fitting the multistate cure model in the presence of covariate missingness and/or unequal follow-up of the two outcomes and we describe a novel approach for obtaining standard errors. Simulations demonstrate good algorithmic performance as long as the modeling assumptions are sufficiently restrictive. We apply the proposed algorithm to a study of head and neck cancer.

✉ lbeesley@umich.edu

PRIORITIZED CONCORDANCE INDEX FOR COMPOSITE SURVIVAL OUTCOMES

Li C. Cheung*, *National Cancer Institute, National Institutes of Health*

Qing Pan, *George Washington University*

Noorie Hyun, *National Cancer Institute, National Institutes of Health*

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

Composite measures are increasingly common in biomedical studies to boost power, reduce cost, or to fully account for multifaceted diseases. Harrell's concordance (C) index, widely used to evaluate predictions from regression models for univariate survival outcomes, does not apply to multivariate survival outcomes. We propose the prioritized concordance index, an extension of the C index that uses the most important comparable outcome for each subject pair. We apply the prioritized concordance index to disease processes with a rare primary outcome and a more common secondary outcome. Asymptotic properties are derived using U-statistic properties. Our simulation studies show that the new concordance index gain efficiency and power in identifying true prognostic variables compared to Harrell's C index for the primary outcome alone. Using the prioritized concordance index, we examine whether novel clinical measures can be useful in predicting risks of type II diabetes in patients with impaired glucose resistance.

✉ li.cheung@nih.gov

MULTI-LEVEL VARIABLE SELECTION FOR MARGINAL PROPORTIONAL HAZARDS MODEL

Natasha A. Sahr*, *Medical College of Wisconsin*

Soyoung Kim, *Medical College of Wisconsin*

Kwang Woo Ahn, *Medical College of Wisconsin*

Variable selection methods for the marginal proportional hazards model is a relatively understudied research area in biostatistics. The limited available methods focus on the selection of non-zero individual variables. However, variable selection in the presence of grouped covariates is often required. Some methods are available for the selection of non-zero group and within-group variables for the univariate proportional hazards model. There are no available methods to perform group variable selection in the clustered multivariate survival setting. In this context, we propose the hierarchical adaptive group bridge penalty to select non-zero group and within-group variables for the marginal proportional hazards model with independent or clustered multivariate failure time data. The simultaneous selection of non-zero group and within-group variables for multivariate modeling is defined as multi-level selection. The simulation studies show that the hierarchical adaptive group bridge method has superior performance compared to the extension of the adaptive group bridge in terms of variable selection accuracy.

✉ nsahr@mcw.edu

53. NONPARAMETRIC METHODS

L-STATISTICS FOR QUANTIFYING THE AGREEMENT BETWEEN TWO VARIABLES

Elahe Tashakor*, *The Pennsylvania State Health Milton S. Hershey Medical Center*

Vernon M. Chinchilli, *The Pennsylvania State Health Milton S. Hershey Medical Center*

The importance of evaluating the agreement between two raters via the Concordance Correlation Coefficient (CCC) has recently received much attention. Since its introduction by Lin (1989), an increasing number of papers proposed to estimate the CCC under a variety of statistical models. Most commonly, it is used under the assumption that data are normally distributed. However, in many practical applications, data are often skewed and/or thick-tailed. Robust estimators of the CCC have been developed by King and Chinchilli (2001) and remain a controversial issue for statistical research. We propose an approach that extends the existing methods of robust estimators of CCC by focusing on functionals that yield robust L-statistics. We provide two data examples to illustrate the methodology, and we discuss the results of computer simulation studies that evaluate statistical performance.

✉ eqt5124@psu.edu

ASSESSING ALIGNMENT BETWEEN FUNCTIONAL MARKERS AND ORDINAL OUTCOMES BASED ON BROAD SENSE AGREEMENT

Jeong Hoon Jang*, *Emory University*

Limin Peng, *Emory University*

Amita K. Manatunga, *Emory University*

The concept of broad sense agreement (BSA) has recently been introduced for studying the relationship between continuous and ordinal measurements (Peng et al., 2011). In this paper, we propose a framework based on BSA that is useful for assessing alignment between a functional marker and an ordinal outcome. We adopt a general class of summary functionals, each of which flexibly captures a different quantitative feature of a functional marker and its (higher-order) derivatives. This approach allows studying alignment between a large class of important features of a functional marker and an ordinal outcome by evaluating BSA based on real-valued outputs of the corresponding summary functionals. The proposed BSA estimator is proven to be consistent and asymptotically normal. We further illustrate the proposed framework using three widely-used classes of summary functionals. In addition, we provide an inferential framework for identifying the summary functional that exhibits better correspondence with the ordinal outcome. Our simulation results demonstrate satisfactory performance of the proposed framework. We demonstrate the application of our methods using a renal study.

✉ jjang54@emory.edu

ON QUANTILES ESTIMATION BASED ON DIFFERENT STRATIFIED SAMPLING WITH OPTIMAL ALLOCATION

Hani M. Samawi*, *Georgia Southern University*

Jingjing Yin, *Georgia Southern University*

Arpita Chatterjee, *Georgia Southern University*

Haresh Rochani, *Georgia Southern University*

In this work, we consider the problem of estimating a quantile function based on different stratified sampling mechanisms. First, we develop an estimate for population quantiles based on stratified simple random sampling (SSRS) and extend the discussion for stratified ranked set sampling (SRSS). We also study the asymptotic behavior of the proposed estimators. Here, we derive an analytical expression for the optimal allocation under both sampling schemes. Simulation studies are designed to examine the performance of the proposed estimators under varying distributional assumptions. The efficiency of the proposed estimates is further illustrated by analyzing a real data set containing TC biomarker values taken from 10,187 Chinese children and adults (>age 7) in the year 2009.

✉ samawi.hani2@gmail.com

ADJUSTED EMPIRICAL LIKELIHOOD BASED CONFIDENCE INTERVAL OF ROC CURVES

Haiyan Su*, *Montclair State University*

We propose an adjusted empirical likelihood (AEL) based confidence interval for receiver operating characteristic curves which are based on a continuous-scale test. The AEL based approach is simply implemented, and computationally efficient. The results from the simulation studies indicate that the finite-sample numerical performance slightly outperforms the exisiting methods. Real data is analyzed by using the proposed method and the existing bootstrap-based method.

✉ suh@mail.montclair.edu

EXACT NONPARAMETRIC CONFIDENCE INTERVALS FOR QUANTILES

Xin Yang*, *State University of New York at Buffalo*

Alan D. Hutson, *Roswell Park Cancer Institute and State University of New York at Buffalo*

Dongliang Wang, *State University of New York Upstate Medical University*

In this article, we develop a kernel-type density estimator for a single order statistic by approximating the convolution of the

kernel and the single order statistic density. The idea is further used to construct the nonparametric confidence interval for an arbitrary quantile based on a Studentized-t analogy, which is distinct from the conventional percentile-t bootstrap method in that it is analytically and computationally feasible to provide an exact estimate of the distribution without resampling. The accuracy of the coverage probabilities is examined via a simulation study. An application to the extreme quantile problem in flood data is illustrated.

✉ xyang.krystal@gmail.com

NONIDENTIFIABILITY IN THE PRESENCE OF FACTORIZATION FOR TRUNCATED DATA

Jing Qian*, *University of Massachusetts, Amherst*

Bella Vakulenko-Lagun, *Harvard School of Public Health*

Sy Han Chiou, *University of Texas, Dallas*

Rebecca A. Betensky, *Harvard School of Public Health*

A time to event, X, is left truncated by T if X can be observed only if T is less than X. This often results in over sampling of large values of X, and necessitates adjustment of estimation procedures to avoid bias. Simple risk-set adjustments can be made to standard risk-set based estimators to accommodate left truncation as long as T and X are "quasi-independent", i.e., independent in the observable region. Through examination of the likelihood function, we derive a weaker factorization condition for the conditional distribution of T given X in the observable region that likewise permits risk-set adjustment for estimation of the distribution of X (but not T). Quasi-independence results when the analogous factorization condition for X given T holds, as well. While we can test for factorization, if the test does not reject, we cannot identify which factorization condition holds, or whether both (i.e., quasi-independence) hold. Importantly, this means that we must ultimately make an unidentifiable assumption in order to estimate the distribution of X based on truncated data. We illustrate these concepts through examples and a simulation study.

✉ qian@schoolph.umass.edu

54. PHARMACOKINETIC/ PHARMACODYNAMICS AND BIOPHARMACEUTICAL RESEARCH

BAYESIAN INFERENCE FROM A NESTED CASE-COHORT DESIGN LINKED WITH A PHARMACOKINETIC MODEL USING BAYESIAN ADDITIVE REGRESSION TREES TO INFER THE PROTECTIVE EFFECT OF TENOFOVIR AGAINST HIV INFECTION

Claire F. Ruberman*, *Johns Hopkins Bloomberg School of Public Health*

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

Gary L. Rosner, *Johns Hopkins School of Medicine*

Craig W. Hendrix, *Johns Hopkins School of Medicine*

Katarina Vucicevic, *University of California, San Francisco*

Rada Savic, *University of California, San Francisco*

Although randomized trials have shown pre-exposure prophylaxis to be highly successful in reducing the risk of HIV infection, much uncertainty remains about the drug concentrations necessary to protect against infection. Key challenges in estimating the protective effect of drug levels in the body include that data on drug concentrations is relatively sparse and is collected via nested case-cohort sampling within the active treatment arm(s), adherence to assigned study drug may vary by study visit, and study visits may be missed. We use a population pharmacokinetic (PK) model developed from the drug concentration data pooled across multiple trials to estimate concentration levels in study participants over time and individual probabilities of treatment compliance at each visit. We then employ Bayesian Additive Regression Trees, based off of output from the PK model, to predict concentration levels for study

participants lacking concentration data. Using the imputed data set of concentration levels for all treated participants, we build a Bayesian hierarchical model to make inferences about the longitudinal relationship between drug exposure and risk of HIV infection.

✉ claireruberman@gmail.com

BAYESIAN PERSONALIZED MULTI-CRITERIA BENEFIT-RISK ASSESSMENT OF MEDICAL PRODUCTS

Kan Li*, *University of Texas Health Science Center at Houston*

Sheng Luo, *Duke University Medical Center*

The evaluation of a medical product always requires a benefit-risk (BR) assessment. To respond to the Patient-Centered Benefit-Risk (PCBR) project commissioned by the US Food and Drug Administration, we propose a Bayesian personalized multicriteria decision-making method for BR assessment. This method is based on a multidimensional latent trait model and a stochastic multicriteria acceptability analysis approach. It can effectively account for the subject-level differences in treatment effects, dependencies among BR criteria, and incorporate imprecise or heterogeneous patient preference information. One important feature of the method is that it focuses on the perspective of patients who live with a disease and are directly impacted by the regulatory decision and treatments. We apply the method to a real example to illustrate how it may improve the transparency and consistency of the decision-making. The proposed method could facilitate communications of treatment decisions between healthcare providers and individual patients based on personalized BR profiles. It could also be an important complement to the PCBR framework to ensure a patient-centric regulatory approval process.

✉ kan.li@uth.tmc.edu

TWO/THREE-STAGE DESIGNS FOR PHASE I DOSE-FINDING

Wenchuan Guo*, *University of California, Riverside*

Bob Zhong, *Johnson & Johnson*

We propose a new two-/three-stage dose-finding designs for Phase 1 clinical trials, where we link the decision rules in the dose-finding process with the conclusions from a hypothesis test. Our method is an extension of traditional "3+3" design to more general "A+B" or "A+B+C" designs, providing statistical explanations using frequentist framework. This method is very flexible that incorporates other interval-based designs decision rules through different parameter settings. We provide the decision table to guide investigators when to decrease, increase or repeat a dose for next cohort of subjects. We conduct simulation experiments to compare the performance of the proposed method with other dose-finding designs. A free open source R package tsdf is available on GitHub. It is dedicated to calculate two- / three-stage designs decision table and perform dose-finding simulations.

✉ wguo1017@gmail.com

NON-INFERIORITY TESTING FOR THREE-ARM TRIALS WITH BINARY OUTCOME: NOVEL FREQUENTIST AND BAYESIAN PROPOSALS

Shrabanti Chowdhury*, *Wayne State University School of Medicine*

Ram C. Tiwari, *U.S. Food and Drug Administration*

Samiran Ghosh, *Wayne State University School of Medicine*

Necessity for improvement in many therapeutic areas are of high priority due to unwarranted variation in restorative treatment, increasing expense of medical care and poor patient outcomes. Although efficacy is the most important evaluating criteria to measure a treatment's beneficial effect, there are several other important factors(e.g. side effects, cost burden, less debilitating etc.), which can permit some

less efficacious treatment options favorable to a subgroup of patients. This leads to non-inferiority (NI) testing. NI trials may or may not include a placebo arm due to ethical reason. However when included, the resulting three-arm trial is more prudent since it requires less stringent assumptions compared to the two-arm placebo-free trial. In this article, we consider both Frequentist and Bayesian procedure for testing NI in the three-arm trial with binary outcomes. Bayesian paradigm provides a natural path to integrate historical and current trials, as well as uses patients/clinicians opinions as prior information via sequential learning. In addition we discuss sample size calculation and draw an interesting connection between the two paradigms.

✉ gg0658@wayne.edu

BAYESIAN INTERVAL-BASED DOSE FINDING DESIGN WITH QUASI-CONTINUOUS TOXICITY MODEL

Dan Zhao*, *University of Illinois, Chicago*

Jian Zhu, *Takeda Pharmaceuticals*

Eric Westin, *ImmunoGen*

Ling Wang, *Takeda Pharmaceuticals*

Current oncology dose-finding designs dichotomize adverse events of various types and grades within the first treatment cycle into binary outcomes (e.g. dose-limiting toxicity). Such inefficient use of information often results in imprecise MTD estimation. To avoid this, Yin et al. (2016) proposed a Bayesian repeated measures design to model a semi-continuous endpoint that incorporates toxicity types and grades from multiple cycles. However, this design follows a decision rule that selects the dose minimizing a point-estimate-based loss function, which can be less reliable due to small sample sizes. To address this concern, we proposed an interval-based design that selects dose with the highest posterior probability of being in a pre-specified target toxicity interval. Through simulation, we compared our design with the original design and popular designs such as the continual reassessment method. The results demonstrated that our design outperforms all other designs in terms of accurately identifying the target dose and assigning more patients to effective dose levels.

✉ danzhao3117@gmail.com

A BAYESIAN FRAMEWORK FOR INDIVIDUALIZING TREATMENT WITH THERAPEUTIC DRUG MONITORING

Hannah L. Weeks*, *Vanderbilt University*

Ryan T. Jarrett, *Vanderbilt University*

William H. Fissell, *Vanderbilt University*

Matthew S. Shotwell, *Vanderbilt University*

Due to dramatic pharmacokinetic heterogeneity, continuous assessment of pharmacodynamic target attainment (PDTA) may be critical for effective antibiotic therapy and mitigation of toxicity risks. Using a Bayesian compartmental model and prior pharmacokinetic data, we developed statistical methodology and a web application that facilitate assessment of individual pharmacokinetics in real time. Application users enter dosing characteristics for a given patient and may update the model with drug concentration measurements, which indicate how patient-specific pharmacokinetics are affecting response to treatment. The application provides an estimate of PDTA with a measure of statistical uncertainty using Laplace and delta method approximations. A tool of this nature allows physicians to tailor dosing to an individual in order to improve the probability of effective and safe treatment. In evaluating our methodology, approximations are slightly anti-conservative. While approximate methods can be used for investigating various infusion schedules, exact intervals obtained via Markov chain Monte Carlo simulation provide accurate interval estimates at the expense of computation time.

✉ hannah.weeks@vanderbilt.edu

55. ORAL POSTERS: MEDICAL IMAGING

55a. INVITED ORAL POSTER: EXPLORATORY TOOLS FOR DYNAMIC CONNECTIVITY AND LOW DIMENSIONAL REPRESENTATIONS OF BRAIN SIGNALS

Hernando Ombao*, *King Abdullah University of Science and Technology*

Hector Flores, *University of California, Irvine*

Abdulrahman Althobaiti, *Rutgers University*

Altyn Zhelambayeva, *Nazarbayev University*

The key challenges to brain signal analysis are the high dimensionality, size of data and complex dependence structures between brain regions. In this poster will present a set of novel exploratory tools that we have developed for creating low-dimensional representations of high dimensional brain signals and for investigating lead-lag dependence between brain signals through their various oscillatory components. We will compare signal summaries obtained from various methods such as spectral principal components analysis and the generalized dynamic principal components analysis. Moreover, different dynamic connectivity measures will be presented: partial coherence, partial directed coherence, evolutionary dual-frequency coherence and lagged dual-frequency coherence. These methods will be illustrated on a variety of brain signals: rat local field potentials recorded during induced stroke and human electroencephalogram in an auditory task.

✉ hernando.ombao@kaust.edu.sa

55b. INVITED ORAL POSTER: REGRESSION MODELS FOR COMPLEX BIOMEDICAL IMAGING DATA

Jeff Morris*, *University of Texas MD Anderson Cancer Center*

Hongxiao Zhu, *Virginia Tech*

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

Hojin Yang, *University of North Carolina, Chapel Hill*

Biomedical imaging produces complex, high dimensional data that poses significant analytical challenges. In practice, many investigators use a feature extraction approach to analyze these data, which involves computing summary measures from the imaging and analyzing those while discarding the original raw image data. If the summary measures contain the meaningful scientific features of the images, this approach can work well, but other times there is important information in the images that are not captured by these features so is lost to analysis. In this presentation, I will present two approaches for analyzing image data that involve flexible modeling that attempts to retain maximal information from the raw data while accounting for their complex structure. First, I will present functional regression methods for modeling event-related potential data that accounts for their complex spatial-temporal correlation structure and identifies regions of the sensor space and time related to factors of interest while accounting for multiple testing. Second, I will present methods to regress the entire marginal distribution of pixel intensities on predictors using a method we call quantile functional regression, which allows us to globally test which covariates affect the distribution, and then determining which distributional features, e.g. which quantiles or moments, characterize these differences. We show that these methods find important biological differences that would have been missed by more naïve simple approaches.

✉ jefmorris@mdanderson.org

55c. INVITED ORAL POSTER: PENALIZED MODELS TO DETECT SUBTLE MULTIPLE SCLEROSIS ABNORMALITIES IN WHITE AND GREY MATTER USING FUNCTIONAL DATA ANALYSIS OF MULTIPLE NON-CONVENTIONAL MRI CONTRASTS

Lynn E. Eberly*, *University of Minnesota*

Kristine Kubisiak, *Chronic Disease Research Group*

Mark Fiecas, *University of Minnesota*

Quantitative methods to detect subtle abnormalities in normal-appearing brain matter in multiple sclerosis (MS) may further our understanding of the pathophysiology and progression of MS. Commonly, voxel level data are summarized to a region of interest (ROI) using a summary statistic such as a mean, but this is likely to be an adequate representation only when the within-ROI distributions are approximately normal with common variance. We use the estimated probability density functions (pdf) of the ROI's voxel-level magnetic resonance (MR) metrics to detect subtle abnormalities in subcortical grey matter (GM) and white matter (WM) of MS patients compared to age-matched controls. A penalized logistic regression model detects MS based on a functional data analysis of how far each individual's pdf is from a 'central' pdf, using thirteen different MR metrics. Compared to using summary statistics, our method detects subtle differences in otherwise-normal-appearing subcortical GM and WM of MS patients with high sensitivity and specificity. This method may provide a more accurate and robust prognostic marker of lesion formation and overall disease progression than conventional methods.

✉ leberly@umn.edu

55d. MIMoSA: A METHOD FOR INTER-MODAL SEGMENTATION ANALYSIS OF T2 HYPERINTENSITIES AND TI BLACK HOLES IN MULTIPLE SCLEROSIS

Alessandra M. Valcarcel*, *University of Pennsylvania*

Kristin A. Linn, *University of Pennsylvania*

Fariha Khalid, *Brigham and Women's Hospital*

Simon N. Vandekar, *University of Pennsylvania*

Theodore D. Satterthwaite, *University of Pennsylvania*

Rohit Bakshi, *Brigham and Women's Hospital*

Russell T. Shinohara, *University of Pennsylvania*

Magnetic resonance imaging (MRI) is crucial for detection and characterization of white matter lesions (WML) in multiple sclerosis. The most widely established MRI outcome measure is T2-weighted lesion (T2L) volume. Unfortunately, T2L volume is non-specific for the level of tissue destruction and shows a weak relationship to clinical status. Consequently, researchers have focused on T1-weighted hypointense lesion (T1L) volume quantification to provide more specificity for axonal loss and a closer link to neurologic disability. This study aimed to adapt and assess the performance of an automatic T2L segmentation algorithm for segmenting T1L. T1, T2, and FLAIR sequences were acquired from 40 MS subjects and with manually segmented T2L and T1L. We employ MIMoSA, an automated segmentation algorithm built to segment T2L. MIMoSA utilizes complementary MRI pulse sequences to emphasize different tissue properties, which can help identify and characterize interrelated features of lesions, in a local logistic regression to model the probability that any voxel is part of a lesion. Using bootstrap cross-validation, we found that MIMoSA is a robust method to segment both T2L and T1L.

✉ alval@pennmedicine.upenn.edu

55e. SPATIALLY ADAPTIVE COLOCALIZATION ANALYSIS IN DUAL-COLOR FLUORESCENCE MICROSCOPY

Shulei Wang*, *University of Wisconsin, Madison and Columbia University*

Ellen T. Arena, *University of Wisconsin, Madison*

Jordan T. Becker, *University of Wisconsin, Madison*

William M. Bement, *University of Wisconsin, Madison*

Nathan M. Sherer, *University of Wisconsin, Madison*

Kevin W. Eliceiri, *University of Wisconsin, Madison*

Ming Yuan, *Columbia University and University of Wisconsin, Madison*

Colocalization analysis aims to study complex spatial associations between bio-molecules via optical imaging techniques. However, existing colocalization analysis workflows only assess an average degree of colocalization within a certain region of interest and ignore the unique

and valuable spatial information offered by microscopy. In the current work, we introduce a new framework for colocalization analysis that allows us to quantify colocalization levels at each individual location and automatically identify spots or regions where colocalization occurs. The framework, referred to as spatially adaptive colocalization analysis (SACA), integrates a pixel-wise local kernel model for colocalization quantification and a multi-scale adaptive propagation-separation strategy for utilizing spatial information to detect colocalization in a spatially adaptive fashion. Applications to simulated and real biological datasets demonstrate the practical merits of SACA in what we hope to be an easily applicable and robust colocalization analysis method. In addition, theoretical properties of SACA are investigated to provide rigorous statistical justification.

✉ shulei.wang364@gmail.com

55f. A LONGITUDINAL MODEL FOR FUNCTIONAL CONNECTIVITY NETWORKS USING RESTING-STATE fMRI

Brian B. Hart*, *University of Minnesota*

Ivor Cribben, *University of Alberta*

Mark Fiecas, *University of Minnesota*

Many studies collect functional magnetic resonance imaging (fMRI) data longitudinally. However, the current literature lacks a general framework for analyzing functional connectivity (FC) networks in longitudinal fMRI data. We build a longitudinal FC network model using a variance components approach. First, for all subjects' visits, we account for the autocorrelation inherent in fMRI time series. Second, we use generalized least squares to estimate 1) the within-subject variance component 2) the FC network, and 3) the FC network's longitudinal trend. Our novel method for longitudinal FC networks accounts for the within-subject dependence across multiple visits, the variability from subject heterogeneity, and the autocorrelation present in fMRI data, while restricting the parameter space to make the method computationally feasible. We develop a permutation testing procedure for valid inference on group differences in baseline FC and longitudinal change in FC between patients and controls. To examine performance, we run a series of simulations and apply the model to longitudinal fMRI data collected from the Alzheimer's Disease Neuroimaging Initiative database.

✉ hartx204@umn.edu

55g. LOW-RANK STRUCTURE BASED BRAIN CONNECTIVITY GWAS STUDY

Ziliang Zhu*, *University of North Carolina, Chapel Hill*

Fan Zhou, *University of North Carolina, Chapel Hill*

Liuqing Yang, *University of North Carolina, Chapel Hill*

Yue Shan, *University of North Carolina, Chapel Hill*

Jingwen Zhang, *University of North Carolina, Chapel Hill*

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

Hongtu Zhu, *University of Texas MD Anderson Cancer Center*

In this paper, we propose a new method for doing connectivity GWAS based on spectral clustering to detect the low-rank structure of brain connectivity and also overcome the drawback of the high dimensionality. In the first step, we perform a spectual clustering algorithm to detect the low rank structure of brain connectivity, and thus extracting only a few features for analysis. The second step is to perform multidimensional phenotype GWAS analysis on the features extracted in the first step.

✉ ziliang@live.unc.edu

55h. BAYESIAN INTEGRATIVE ANALYSIS OF RADIOGENOMICS

Youyi Zhang*, *University of Texas MD Anderson Cancer Center*

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

Shivali Narang Aerry, *Johns Hopkins University*

Arvind U.K. Rao, *University of Texas MD Anderson Cancer Center*

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

We present a multi-stage integrative Bayesian hierarchical model for the analysis of Radiogenomics (imaging genetics) driven by the motivation of linking non-invasive imaging features, multiplatform genomics information and clinical outcomes. Our goals are to identify significant genes and imaging markers as well as the hidden associations between these two platforms, and to further detect the overall clinical relevance. For this task, we established a multi-stage Bayesian hierarchical model which acquires several annovative characteristics: it incorporates integrative analysis of multi-platform genomics data sets to capture fundamental biological mechanism in Radiogenomics framework; explores the associations between imaging markers carrying genetic information with clinical outcomes; detects important genetic markers and imaging markers via establishing hierarchical model with Bayesian continuous shrinkage priors. Applied to the Glioblastoma (GBM) dataset, the model hierarchically identifies important magnetic resonance imaging (MRI) imaging features and the associated genomic platforms that significantly affect patients' survival.

✉ youyimimi66@gmail.com

55i. HOW TO EXPLOIT THE BRAIN CONNECTIVITY INFORMATION AND INCREASE THE ESTIMATION ACCURACY UNDER REPEATED MEASURES DESIGN?

Damian Brzyski*, *Indiana University, Bloomington*

Marta Karas, *Johns Hopkins University*

Beau Ances, *Washington University School of Medicine*

Joaquin Goni, *Purdue University*

Timothy W. Randolph, *Fred Hutchinson Cancer Research Center*

Jaroslaw Harezlak, *Indiana University, Bloomington*

One of the challenging problems in the brain imaging research is a principled incorporation of information from different imaging modalities in association studies. Often, data from each modality is analyzed separately using, for instance, dimensionality reduction techniques, which results in a loss of mutual information. Another important problem to address is the incorporation of correlations among observations arising from repeated measurements on the same subject in a longitudinal study. We propose a novel regularization method, rePEER (repeated Partially Empirical Eigenvectors for Regression) to estimate the association between the brain structure features and a scalar outcome. Our approach employs the external information about the brain connectivity and takes into account the repeated measures designs. The method we propose is formulated as a penalized convex optimization problem. We address theoretical and computational issues, such as the selection of tuning parameters. We evaluated the performance of rePEER in simulation studies and applied it to analyze the association between cortical thickness and HIV-related outcomes in the group of HIV-positive individuals.

✉ dbrzyski@iu.edu

55j. ASSESSING THE RELATIONSHIP BETWEEN CORTICAL THINNING AND MYELIN MEASUREMENTS IN MULTIPLE SCLEROSIS DISEASE SUBTYPES: A WHOLE BRAIN APPROACH

Sandra Hurtado Rua*, *Cleveland State University*

Michael Dayan, *Weill Cornell Medicine*

Susan A. Gauthier, *Weill Cornell Medicine*

Elizabeth Monohan, *Weill Cornell Medicine*

Kyoko Fujimoto, *Weill Cornell Medicine*

Sneha Pandya, *Weill Cornell Medicine*

Eve LoCastro, *Weill Cornell Medicine*

Tim Vartanian, *Weill Cornell Medicine*

Thanh D. Nguyen, *Weill Cornell Medicine*

A lesion-mask free method based on a gamma mixture (GM) model was applied to MRI myelin water fraction (MWF) maps and the association between cortical thickness and myelin was estimated for relapsing-remitting (RRMS) and secondary-progressive multiple sclerosis (SPMS) patients. The GM model of whole brain white matter (WM) MWF was characterized with three variables: The mode (most frequent value) of the gamma first component shown to relate to lesion, the mode of the second component shown to be associated with normal appearing WM, and the mixing ratio (?) between the two distributions. A regression analysis was carried out to find the best predictors of cortical thickness for each group. The results suggest that during the relapsing phase, focal WM damage is associated with cortical thinning, yet in SPMS patients, global WM deterioration has a much stronger influence on secondary degeneration. We demonstrate the potential contribution of myelin loss on neuronal degeneration at different disease stages and the usefulness of the statistical reduction technique.

✉ s.hurtadorua@csuohio.edu

55k. BAYESIAN JOINT MODELING OF MULTIPLE BRAIN FUNCTIONAL NETWORKS

Joshua Lukemire*, *Emory University*

Suprateek Kundu, *Emory University*

Giuseppe Pagnoni, *University of Modena and Reggio Emilia*

Ying Guo, *Emory University*

Brain function is organized in coordinated modes of spatio-temporal activity (functional networks) exhibiting an intrinsic baseline structure with variations under different experimental conditions. Existing approaches for uncovering such network structures typically do not explicitly model shared and differential patterns across networks, thus potentially reducing the detection power. We develop an integrative modeling approach for jointly modeling multiple brain networks across experimental conditions. The proposed Bayesian Joint Network Learning approach develops flexible priors on the edge probabilities involving a common intrinsic baseline structure and differential effects specific to individual networks. Conditional on these edge probabilities, connection strengths are modeled under a Bayesian spike and slab prior on the off-diagonal elements of the inverse covariance matrix. The model is fit under a posterior computation scheme based on Markov chain Monte Carlo. An application of the method to fMRI Stroop task data provides unique insights into brain network alterations between cognitive conditions.

✉ joshua.lukemire@emory.edu

56. QUANTIFYING COMPLEX DEPENDENCY

DEPENDENCE MEASURES: SOMETHING OLD AND SOMETHING NEW, SOMETHING BORROWED, AND SOMETHING BLUE

Gabor J. Szekely*, *National Science Foundation*

Starting with Francis Galton (1888) and Karl Pearson (1896) many researchers have introduced dependence measures in the past 130 years. Distance correlation was introduced by the speaker in 2005. In this talk we propose four simple axioms for dependence measures and then discuss the "Theorem in Blue" that most of the frequently applied dependence measures fail to satisfy these axioms. For example the empirical maximal correlation is always 1 even if the underlying variables are independent. The same can happen with the recently introduced maximal information coefficient. From this point of view distance correlation is a good candidate for an ideal dependence measure for the 21st century because distance correlation is continuous and thus robust, it is zero if and only if the variables are

independent and distance correlation is also invariant with respect to all similarity transformations. Affine invariance would contradict to continuity.

✉ gszekely@nsf.gov

BET ON INDEPENDENCE

Kai Zhang*, *University of North Carolina, Chapel Hill*

We study the problem of nonparametric dependence detection in copula. Many existing methods suffer severe power loss due to non-uniform consistency, which we illustrate with a paradox. To avoid such power loss, we approach the nonparametric test of independence through a novel binary expansion filtration approximation. Through a Hadamard-Walsh transform, we show that the cross interactions of binary variables in the filtration are complete sufficient statistics for dependence. These interactions are also uncorrelated under the null. By utilizing these interactions, the resulting method of binary expansion testing (BET) avoids the problem of non-uniform consistency and improves upon a wide class of commonly used methods (a) by achieving the optimal rate in sample complexity and (b) by providing clear interpretations of global and local relationships upon rejection of independence. The binary expansion approach also connects the test statistics with the current computing system to facilitate efficient bitwise implementation. We illustrate the BET by an exploratory data analysis of the TCGA breast cancer data.

✉ zhangk@email.unc.edu

FISHER EXACT SCANNING FOR DEPENDENCY

Li Ma*, *Duke University*

Jialiang Mao*, Duke University*

We introduce Fisher exact scanning (FES) for testing and identifying variable dependency. FES proceeds through scanning over the sample space using windows in the form of 2 by 2 tables of various sizes, and on each window completing a Fisher exact test. Based on a factorization of multivariate hypergeometric (MHG) likelihood into the product of univariate hypergeometric likelihoods, we show that there exists a coarse-to-fine, sequential generative representation for the MHG model in the form of a Bayesian network, which in turn implies the mutual independence (up to deviation due to discreteness) among the Fisher exact tests completed under FES. This allows exact characterization of the joint null distribution of the p-values and gives rise to an effective inference recipe through simple multiple testing procedures such as Sidak and Bonferroni corrections, eliminating the need for resampling. FES can characterize dependency through reporting significant windows after multiple testing control. The computational complexity of FES is approximately linear in the sample size, which along with the avoidance of resampling makes it ideal for analyzing massive data sets.

✉ li.ma@duke.edu

GENERALIZED R-SQUARED FOR MEASURING DEPENDENCE

Jun Liu*, *Harvard University*

Xufei Wang*, Two Sigma Inc.*

Bo Jiang, *Two Sigma Inc.*

Detecting and quantifying dependence between two random variables is a fundamental problem. Although the Pearson correlation is effective for capturing linear dependency, it can be entirely powerless for detecting nonlinear and/or heteroscedastic patterns. We introduce a new measure, G-squared, to measure how much two random variables are related and test whether they are independent. The G-squared is almost identical to the square of the classic R-squared for linear relationships with constant error variance, and has the intuitive meaning of the piecewise R-squared between the variables. It is particularly effective in handling nonlinearity and heteroscedastic errors. We propose two estimators of G-squared and show their consistency. Simulations demonstrate that G-squared estimators are among the most powerful test statistics compared with several state-of-the-art methods.

✉ jliu@stat.harvard.edu

57. PREPARING FOR THE JOB MARKET

PANEL DISCUSSANTS:

Pallavi Mishra-Kalyani, *U.S. Food and Drug Administration*

Brooke Alhanti, *North Carolina State University*

Barbara Wendelberger, *Berry Consultants*

Ning Leng, *Genentech*

58. NOVEL CLINICAL TRIAL DESIGNS

NOVEL RESPONSE ADAPTIVE ALLOCATIONS IN FACTORIAL DESIGNS: A CASE STUDY

John A. Kairalla*, *University of Florida*

Rachel S. Zahigian, *University of Florida*

Samuel S. Wu, *University of Florida*

Response adaptive randomization uses observed treatment outcomes from preceding participants to change allocation probabilities. Traditionally, the strategy can fulfill the ethical desire to increase the likelihood of giving an individual the best-known treatment at the time of randomization. In a multi-arm clinical trial setting with ordered testing priorities, novel response adaptive allocation methods may allow for more flexibility and efficiency with respect to information allocation decisions made during study accrual. We will review two such novel response adaptive allocation designs recently funded by the NIH that are currently in early accrual phases. Both are multi-stage 2x2 factorial designs with fixed total sample size. In one, studying biopsychosocial influence on shoulder pain, the primary hypothesis is tested at both the interim and final stages. The other, studying augmented cognitive training in older adults, involves interim testing for a secondary hypothesis to go along with allocation decisions. Study operating characteristics will be extensively explored, summarized, and compared to alternatives, with recommendations for improvements to the designs given.

✉ jak@biostat.ufl.edu

METHODS AND SOFTWARE FOR OPTIMIZING ADAPTIVE ENRICHMENT DESIGNS

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

Jon Arni Steingrimsson, *Brown School of Public Health*

Josh Betz, *Johns Hopkins Bloomberg School of Public Health*

Aaron Joel Fisher, *Harvard School of Public Health*

Tianchen Qian, *Harvard University*

Adi Gherman, *Johns Hopkins Bloomberg School of Public Health*

Yu Du, *Johns Hopkins Bloomberg School of Public Health*

Adaptive enrichment designs involve preplanned rules for modifying patient enrollment criteria based on data accrued in an ongoing trial. These designs may be useful when it is suspected that a subpopulation, e.g., defined by a biomarker or risk score measured at baseline, may benefit more from treatment than the complementary subpopulation. Our contribution is a new class of adaptive enrichment designs and an open-source software tool that optimizes such designs for a given trial context. We present case-studies showing the potential advantages and limitations of such designs in simluation studies based on data from completed trials involving stroke, HIV, heart failure, and Alzheimer's disease. The adaptive designs are compared to standard designs in terms of the following performance criteria: power, Type I error, sample size, duration, estimator bias and variance, confidence interval coverage probability, and the number of trial participants assigned to an inferior treatment.

✉ mrosen@jhu.edu

THE ADAPTIVE LEARN-AS-YOU-GO DESIGN FOR MULTI-STAGE INTERVENTION STUDIES

Judith J. Lok*, *Harvard School of Public Health*

Daniel Nevo, *Harvard School of Public Health*

Donna Spiegelman, *Harvard School of Public Health*

In learn-as-you-go studies, the intervention is a package consisting of one or more components, and is changed over time and adapted based on past outcome results. This regularly happens in public health intervention studies. The main complication in the analysis is that the interventions in the later stages depend on the outcomes in the previous stages. Therefore, conditioning on the interventions would lead to effectively conditioning on the earlier-stages outcomes, which violates common statistical principles. We have developed a method to estimate treatment effects from a learn-as-you-go study. Our method is based on maximum likelihood estimation. We prove consistency and asymptotic normality using a coupling argument. Typically, one would want to have good efficacy of the intervention package, with limited cost. This leads to a restricted optimization problem with estimated parameters plugged-in. A simulation study indicates that our method works well already in relatively small samples. Moreover, we will present an application to the BetterBirth Study, which aims to increase the use of a checklist when women give birth, in order to improve maternal and fetal health in India.

✉ jlok@hsph.harvard.edu

59. NEW METHODS IN BRAIN CONNECTIVITY

BAYESIAN LOW-RANK GRAPH REGRESSION MODELS FOR MAPPING HUMAN CONNECTOME DATA

Eunjee Lee*, *University of Michigan*

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

Yong Fan, *University of Pennsylvania*

Hongtu Zhu, *University of North Carolina, Chapel Hill and University of Texas MD Anderson Cancer Center*

We propose a Bayesian low-rank graph regression modeling (BLGRM) framework for the regression analysis of matrix response data across subjects. This development is motivated by performing comparisons of functional connectivity data across subjects, groups, and time and relating connections to particular behavioral measures. The BLGRM can be regarded as a novel integration of principal component analysis, tensor decomposition, and regression models. In BLGRM, we find a common low-dimensional subspace for efficiently representing all matrix responses. Based on such low-dimensional representation, we can quantify the effects of various predictors of interest and then perform regression analysis in the common subspace, leading to both dimension reduction and much better prediction. We adapt a parameter expansion approach to our graph regression model (PX-BLGRM) to address weak identifiability and high posterior dependence among parameters in our model. A simulation study is performed to evaluate the performance of BLGRM and its comparison with several competing approaches. We apply BLGRM to the resting-state fMRI data set obtained from the ADNI study.

✉ eunjee@umich.edu

METHODS FOR LONGITUDINAL COMPLEX NETWORK ANALYSIS IN NEUROSCIENCE

Heather Shappell*, *Johns Hopkins Bloomberg School of Public Health*

Yorghos Tripodis, *Boston University*

Ronald J. Killiany, *Boston University*

Eric D. Kolaczyk, *Boston University*

The study of complex brain networks, where the brain can be viewed as a system of interacting regions that produce complex behaviors, has grown notably over the past decade. With an increase in longitudinal study designs and increased interest in the neurological network changes that occur during the progression of a disease, sophisticated methods for dynamic brain network analysis are needed. We propose a paradigm for longitudinal brain network analysis over patient cohorts, where we model a subject's brain network over time as observations of a continuous-time Markov

chain on network space. Network dynamics are represented by various factors, both endogenous (i.e., network effects) and exogenous, which includes mechanisms conjectured in the literature. We outline an application to the resting-state fMRI network setting and demonstrate its use with data from the Alzheimer's Disease Neuroimaging Initiative Study. We draw conclusions at the subject level and compare elderly controls to individuals with AD. Lastly, we extend the models, proposing an approach based on Hidden Markov Models to incorporate and estimate type I and type II error in our observed networks.

✉ hshappe1@jh.edu

CAUSAL MEDIATION ANALYSIS IN NEUROIMAGING

Yi Zhao*, *Johns Hopkins University*

Xi Luo, *Brown University*

Martin Lindquist, *Johns Hopkins University*

Brian Caffo, *Johns Hopkins University*

Causal mediation analysis is widely applied to assess the causal mechanism among three variables: a treatment, an intermediate (i.e., a mediator), and an outcome variable. In neuroimaging studies, neuroscientists are interested in identifying the brain regions that are responsive to an external stimulus, as well as in discovering the pathways that are involved in processing the signals. Functional magnetic resonance imaging (fMRI) is often used to infer brain connectivity, however, mechanistic analysis is challenging given the hierarchically nested data structure, the great number of functional brain regions, and the complexity of data output in the form of times series or functional data. Causal mediation methods in big data contexts are scarce. In this presentation, we will discuss some novel causal mediation approaches aiming to address this methodological gap.

✉ zhaoyi1026@gmail.com

TEMPLATE ICA: ESTIMATING RESTING-STATE NETWORKS FROM FMRI IN INDIVIDUAL SUBJECTS USING EMPIRICAL POPULATION PRIORS

Amanda F. Mejia*, *Indiana University*

Yikai Wang, *Emory University*

Brian Caffo, *Johns Hopkins University*

Ying Guo, *Emory University*

Independent component analysis (ICA) is commonly applied to fMRI data to identify resting-state networks (RSNs), regions of the brain that activate together spontaneously. Due to high noise levels in fMRI, group-level RSNs are typically estimated by combining data from many subjects in a group ICA (GICA). Subject-level RSNs are then estimated by relating GICA results to subject-level fMRI data. Recently, model-based methods that estimate subject-level and group RSNs simultaneously have been shown to result in more reliable subject-level RSNs. However, this approach is computationally demanding and inappropriate for small group or single-subject studies. To address these issues, we propose a model-based approach to estimate RSNs in a single subject using empirical population priors based on large fMRI datasets. We develop an expectation-maximization (EM) algorithm to obtain posterior means and variances of subject-level RSNs. We apply the proposed methods to data from the Human Connectome Project and find that the resulting subject-level RSN estimates are significantly more reliable than those produced from competing methods.

✉ mandy.mejia@gmail.com

60. STATISTICAL METHODS FOR EMERGING SPATIAL AND SPATIOTEMPORAL DATA

A CAUSAL INFERENCE ANALYSIS OF THE EFFECT OF WILDLAND FIRE SMOKE ON AMBIENT AIR POLLUTION LEVELS

Brian J. Reich*, *North Carolina State University*

Alexandra Larsen, *North Carolina State University*

Ana Rappold, *U.S. Environmental Protection Agency*

Wildfire smoke is a major contributor to ambient air pollution levels. In this talk, we develop a spatio-temporal model to estimate the contribution of fire smoke to overall air pollution in different regions of the country. We combine numerical model output with observational data within a causal inference framework. Our methods account for aggregation and potential bias of the numerical model simulation, and address uncertainty in the causal estimates. We apply the proposed method to estimation of ozone and fine particulate matter from wildland fires and the impact on health burden assessment.

✉ brian_reich@ncsu.edu

ADOLESCENT ACTIVITY PATTERNS AND ECOLOGICAL NETWORKS

Catherine Calder*, *The Ohio State University*

Christopher Browning, *The Ohio State University*

Beth Boettner, *The Ohio State University*

Wenna Xi, *The Ohio State University*

Research on neighborhood effects often focuses on linking features of social contexts or exposures to health, educational, and criminological outcomes. Traditionally, individuals are assigned a specific neighborhood, frequently operationalized by the census tract of residence, which may not contain the locations of routine activities. In order to better characterize the many social contexts to which individuals are exposed as a result of the spatially- and temporally-distributed locations of their routine activities and to understand the consequences of these socio-spatial exposures, we have developed the concept of ecological networks. Ecological networks are two-mode networks that indirectly link individuals through the spatial overlap in their routine activities. This presentation focuses on statistical methodology for understanding the structure underlying ecological networks. In particular, we propose a Bayesian mixed-effects models that allows for third-order dependence patterns in the interactions between individuals and the places they visit. We illustrate our methodology using activity pattern and sample survey data from Columbus, OH.

✉ calder@stat.osu.edu

DIAGNOSING GLAUCOMA PROGRESSION WITH VISUAL FIELD DATA USING A SPATIOTEMPORAL BOUNDARY DETECTION METHOD

Joshua L. Warren*, *Yale School of Public Health*

Samuel I. Berchuck, *University of North Carolina, Chapel Hill*

Jean-Claude Mwanza, *University of North Carolina, Chapel Hill*

Diagnosing glaucoma progression early is critical for limiting irreversible vision loss. A common method for assessing glaucoma progression relies on a longitudinal series of visual fields (VF) acquired from a patient at regular intervals. VF data are characterized by a complex spatiotemporal correlation structure due to the data generating process and ocular anatomy. Thus, advanced statistical methods are needed to make clinical determinations regarding progression status. We introduce a spatiotemporal boundary detection model that allows the underlying anatomy of the optic disc to define the spatial structure of the VF data across time. Based on this model, we define a diagnostic metric and verify that it explains a novel pathway in glaucoma progression using data

from the Vein Pulsation Study Trial in Glaucoma and the Lions Eye Institute trial registry. Simulations are presented, showing that the proposed methodology is preferred over an existing spatial boundary detection method for estimation of the new diagnostic measure.

✉ joshua.warren@yale.edu

ON NEW CLASSES OF SPATIAL DISEASE MAPPING MODELS BASED UPON DIRECTED ACYCLIC GRAPHS

Sudipto Banerjee*, *University of California, Los Angeles*

Abhirup Datta, *Johns Hopkins University*

James S. Hodges, *University of Minnesota*

Hierarchical models for regionally aggregated disease incidence data commonly involve region specific latent random effects which are modeled jointly with multivariate Normal distributions. Common choices for the precision matrix include the widely used intrinsic conditional autoregressive model which is singular, and its nonsingular extension which lacks interpretation. We propose a new parametric model for the precision matrix based on a directed acyclic graph (DAG) to introduce spatial dependence. Theoretical and empirical results demonstrate the interpretation of parameters in our model. Our precision matrix is sparse and the model is highly scalable for large datasets. We also derive a novel order-free version which averages over all possible orderings of the DAG. The resulting precision matrix is still sparse and available in closed form. We demonstrate the superior performance of our models over competing models using simulation experiments and a public health application.

✉ sudipto@ucla.edu

6I. NOVEL STATISTICAL LEARNING METHODOLOGIES FOR PRECISION MEDICINE

COMPUTATIONALLY EFFICIENT LEARNING FOR OPTIMAL INDIVIDUALIZED TREATMENT RULES WITH MULTIPLE TREATMENTS

Donglin Zeng*, *University of North Carolina, Chapel Hill*

Xuan Zhou, *University of North Carolina, Chapel Hill*

Yuanjia Wang, *Columbia University*

Powerful machine learning methods have been proposed to estimate an optimal individualized treatment rule, but they are mostly limited to compare only two treatments. When many treatment options are available, which is often the case in practice, how to adapt binary treatment selection rules into a single decision rule is challenging. It is well known in the multicategory learning literature that some approaches may lead to inconsistent decision rules, while the others solve non-convex optimization problems so are computationally intensive. In this work, we propose a novel and efficient method to generalize outcome weighted learning to multi-treatment settings via sequential weighted support vector machines. The proposed method always solves convex optimization problems and computation can be parallelized. Theoretically, we show that the resulting treatment rule is Fisher consistent. Furthermore, we obtain the convergence rate of the estimated value function from the optimal value. We conduct extensive simulations to demonstrate that the proposed method has superior performance to competing methods.

✉ dzeng@email.unc.edu

TREE-BASED REINFORCEMENT LEARNING FOR ESTIMATING OPTIMAL DYNAMIC TREATMENT REGIMES

Lu Wang*, *University of Michigan*

Yebin Tao, *University of Michigan*

Danny Almirall, *University of Michigan*

Dynamic treatment regimes (DTRs) are sequences of treatment decision rules, in which treatment may be adapted over time in response to the changing course of an individual. Motivated by the substance use disorder (SUD) study, we propose a tree-based reinforcement learning (T-RL) method to directly estimate optimal DTRs in a multi-stage multi-treatment setting. At each stage, T-RL builds an unsupervised decision tree that handles the problem of optimization with multiple treatment comparisons directly, through a purity measure constructed with augmented inverse probability weighted estimators. For the multiple stages, the algorithm is implemented recursively using backward induction. By combining robust semiparametric regression with flexible tree-based learning, T-RL is robust, efficient and easy to interpret for the identification of optimal DTRs, as shown in the simulation studies. With the proposed method, we identify dynamic SUD treatment regimes for adolescents.

✉ luwang@umich.edu

EFFECT HETEROGENEITY AND SUBGROUP IDENTIFICATION FOR LONG-TERM INTERVENTIONS

Menggang Yu*, *University of Wisconsin, Madison*

There has been great interest in developing interventions to effectively coordinate the typically fragmented care of patients with many comorbidities. Evaluation of such interventions is often challenging given their long-term nature and their differential effectiveness among diverse patient populations. Given this and the resource intensiveness of care coordination interventions, there is significant interest in identifying which patients may benefit the most from care coordination. We accomplish such goal by modeling covariates which modify the intervention effect. In particular, we consider long-term interventions whose effects are expected to change smoothly over time. We allow interaction effects to vary over time and encourage these effects to be more similar over time by utilizing a fused lasso penalty. Our approach allows for flexibility in modeling temporal effects while also borrowing strength in estimating these effects over time. We use our approach to identify a subgroup of patients who benefit from a complex case management intervention in a large hospital system.

✉ meyu@biostat.wisc.edu

SHARED-PARAMETER G-ESTIMATION OF OPTIMAL TREATMENTS FOR RHEUMATOID ARTHRITIS

Erica E. M. Moodie*, *McGill University*

The doubly-robust method of G-estimation can be used to estimate used an adaptive treatment strategy in which parameters are shared across different stages of the treatment sequence, allowing for more efficient estimation and simpler treatment decision rules. The approach is computationally stable, and produces consistent estimators provided either the outcome model or the treatment allocation model is correctly specified. In this talk, the method will be demonstrated in the context of the treatment of rheumatoid arthritis, a chronic inflammatory condition which can require ongoing treatment.

✉ erica.moodie@mcgill.ca

62. COMPARATIVE EFFECTIVENESS RESEARCH

OPTIMAL WEIGHTS FOR PROPENSITY SCORE STRATIFICATION

Roland A. Matsouaka*, *Duke University*

Propensity score stratification is one of the methods used to control for confounding and reduce bias in the assessment of causal treatment effects. Subjects are group into strata based on their propensity score values, stratum-specific treatment effects are estimated and aggregated into a weighted average treatment effect estimate, where the weights are equal to the proportion of subjects in each stratum. However, these weights are optimal only if the strata are independent and the treatment effect is constant across strata, which is not always true. For this presentation, we first introduce an alternative propensity score stratification approach using weights that maximize the signal-to-noise ratio. Using simulations, we assess the performance of these weights under different data-generating scenarios: vary the number of strata, the propensity score overlap between the treatment groups, and treatment effect across strata. We illustrate the proposed method using data from a cardiovascular disease study.

✉ roland.matsouaka@duke.edu

VARIANCE ESTIMATION FOR THE MATCHED WIN RATIO

Adrian Coles*, *Duke Clinical Research Institute*

Roland A. Matsouaka, *Duke Clinical Research Institute*

The use of composite endpoints in clinical trials has increased in recent years, particularly in cardiovascular trials. Analyzing such composites using a time to first event strategy is problematic as they tend to prioritize less severe components of the composite. The win ratio and the proportion in favor of treatment have been proposed as alternative strategies that allow the prioritization of more severe components of the composite. When estimated from matched data, inference on the win ratio is based on a normal approximation of the binomial distribution, which is only possible when the total number of wins and losses in a treatment group is fixed. We propose large and small sample approaches to estimate confidence intervals for these two quantities from paired samples that do not condition on the total number of wins and losses. We show via simulations that both approches perform well, and we apply our estimators to two recently published heart failure trials.

✉ adrian.coles@duke.edu

CLINICAL TRIAL SIMULATION USING ELECTRONIC MEDICAL RECORDS

Xiaochen Wang*, *Yale University*

Lauren Cain, *Takeda Pharmaceuticals*

Ray Liu, *Takeda Pharmaceuticals*

Dorothy Romanus, *Takeda Pharmaceuticals*

Greg Hather, *Takeda Pharmaceuticals*

Existing clinical trial simulation software is mostly model-based, where parameters are extracted from published trials. However, that approach does not account for differences in enrollment criteria and associations between covariates and outcomes. To produce more realistic trial simulations, we propose a data-based simulation method using electronic medical records (EMR). In our method, outcomes were simulated according to user-supplied trial specifications. Survival times were simulated using a Cox-proportional hazards model to incorporate patients' baseline information. To validate, we simulated the outcomes for patients with newly diagnosed multiple myeloma and compared our results with those of the SWOG S0777 trial. Given differences between the distribution of baseline covariates in EMR and in SWOG, we used weighted sampling where weights were calculated from maximizing empirical likelihood. Median overall survival (OS: 63.9 months, 53.7-Not Estimable) and hazard ratio (HR: 0.604, 0.444-0.832) in our simulation were similar to SWOG (OS: 64 months, 56-Not Estimable; HR: 0.709, 0.524-0.959). More validation results will be shown.

✉ xiaochen.wang@yale.edu

ESTIMATING POPULATION TREATMENT EFFECTS USING META-ANALYSIS

Hwanhee Hong*, *Johns Hopkins Bloomberg School of Public Health*

Elizabeth A. Stuart, Johns Hopkins Bloomberg School of Public Health

Comparative effectiveness research relies heavily on the results of randomized controlled trials (RCTs) to evaluate the efficacy and safety of interventions and inform policy decisions. However, the results of these studies may not generalize to all people in a target population of interest in which we want to make decisions regarding health policy or treatment implementation, because these studies may not have enrolled subjects representative of the target population. Meta-analysis with RCTs is commonly used to evaluate treatments and inform policy decisions because it provides the best summaries of all available evidence. However, meta-analyses are limited to draw population inference of treatment effects because they usually do not define target populations of interest specifically and results of the individual RCTs in those meta-analyses may not generalize to target populations. We extend generalizability methods for a single RCT to meta-analysis with individual participant-level data. We apply these methods to generalize meta-analysis results from RCTs of treatments on schizophrenia to adults with schizophrenia who present to usual care settings in the US.

✉ hhong@jhu.edu

EFFICIENT AND ROBUST SEMI-SUPERVISED ESTIMATION OF AVERAGE TREATMENT EFFECTS IN ELECTRONIC MEDICAL RECORDS DATA

David Cheng* •, *Harvard School of Public Health*

Ashwin Ananthakrishnan, *Massachusetts General Hospital*

Tianxi Cai, *Harvard School of Public Health*

There is strong interest in conducting comparative effectiveness research (CER) in electronic medical records (EMR). However, inferring causal effects in EMR data is challenging due to the lack of direct observation on pre-specified true outcomes. Ascertaining true outcomes often requires labor-intensive medical chart review. Alternatively, average treatment effect (ATE) estimators based on imputations could be biased if the imputation model is mis-specified. We frame ATE estimation in a semi-supervised learning setting, where a small fraction of all observations are labeled with the outcome. We develop an imputation-based approach for estimating the ATE that is robust to mis-specification of the imputation model. The ATE estimator is doubly-robust in that it is consistent under correct specification of either a propensity score or baseline outcome model and locally semiparametric efficient in an ideal semi-supervised model where the distribution of unlabeled data is known. Simulations exhibit the efficiency and robustness of the proposed estimator. We illustrate the method in an EMR study comparing treatment response to two biologic agents for treating inflammatory bowel disease.

✉ dcheng01@fas.harvard.edu

APPLICATIONS OF MULTIPLE IMPUTATION IN THE CONTEXT OF PROPENSITY SCORE MATCHING

Albee Ling*, *Stanford University*

Maya Mathur, *Stanford University*

Kris Kapphahn, *Stanford University*

Maria Montez-Rath, *Stanford University*

Manisha Desai, *Stanford University*

Propensity score (PS) strategies are common for mitigating bias in comparative effectiveness research using observational data. Missing data on key variables used to estimate the PS, however, poses an issue. Including only variables with complete data in the PS models or conducting complete case analysis can lead to biased and inefficient estimates of treatment effects. Multiple Imputation (MI) is a well-established statistical technique under a reasonably flexible set of assumptions. There is no consensus

on best statistical practices for utilizing MI for estimating and integrating PS in the presence of missing data when multiple covariates are missing. We conducted an extensive simulation study to evaluate statistical properties of relevant estimators under a variety of imputation strategies that fall under two umbrellas of MI (MI-passive and MI-active) and that are coupled with two general strategies for integrating PS into analyses (PSI-Within and PSI-Across). We illustrate considerable heterogeneity across approaches in a real study of breast cancer and provide practical guidelines based on findings from our simulation study.

✉ yling@stanford.edu

63. COMPETING RISKS

MODELING OF EXPOSURE-TIME-RESPONSE ASSOCIATION IN THE PRESENCE OF COMPETING RISKS

Xingyuan Li*, *University of Pittsburgh*

Chung-Chou H. Chang, *University of Pittsburgh*

In biomedical studies with long-term follow-up, exposures are often measured over a period of time and have a protracted effect on survival outcome. Also, the intensity of exposure varies, creating challenges to modeling simultaneously the exposure-response association and the time structure since exposure. Meanwhile, an increasing number of clinical studies are involving competing risks where subjects may fail from one of the multiple mutually exclusive events. In this study, we proposed a semiparametric subdistributional hazards regression model to quantify the exposure-time-response association in which the intensity, duration, and timing of an exposure during the study vary among individuals while the event of interest is subject to competing risks. We first defined a weighted time-varying metric to quantify the cumulative effects of an exposure on the event then incorporate cubic B-spline into the partial likelihood equation to estimate the weights. This methodology is demonstrated with an application in Medicare data to investigate the effect of different opioid use patterns on the risk of future opioid overdose, when mortality is treated as a competing event.

✉ xingyuanli96@gmail.com

ANALYSIS OF THE TIME-VARYING COX MODEL FOR CAUSE-SPECIFIC HAZARD FUNCTIONS WITH MISSING CAUSES

Fei Heng, University of North Carolina, Charlotte*

Seunggeun Hyun, University of South Carolina Upstate

Yanqing Sun, University of North Carolina, Charlotte

Peter B. Gilbert, Fred Hutchinson Cancer Research Center

This paper studies the Cox model with time-varying coefficients for cause-specific hazard functions when causes of failure are subject to missingness. This research was motivated by the application to evaluate time-varying cause-specific vaccine efficacy. The inverse probability weighted estimator and augmented inverse probability weighted estimator are investigated. Simulation studies show that the two-stage estimation is more efficient and robust. The proposed methods are illustrated using the Mashi trial data for investigating the effect of formula-feeding versus breast-feeding plus extended infant zidovudine prophylaxis on death due to mother-to-child HIV transmission in Botswana.

✉ fheng@uncc.edu

JOINT RISK PREDICTION IN THE SEMI-COMPETING RISKS SETTING

Catherine Lee*, *Kaiser Permanente Division of Research*

Sebastien Haneuse, *Harvard School of Public Health*

Semicompeting risks refers to the setting where interest lies in the time-to-event for some nonterminal event, the observation of which is subject to some terminal event. We consider prediction in this setting through the calculation and evaluation of patient-specific risk profiles for both events

simultaneously. In particular, at any given point in time after the initiating event, a patient will have experienced: both events; one event without the other; or neither event. In the multi-state model literature, such a profile is derived through the estimation of transition probabilities. We build on that work in two important ways. First, we permit the inclusion of a subject-specific frailty. Second, we consider the evaluation of the predictive performance of the profiles based on the hypervolume under the manifold (HUM) statistic, an extension of the well-known area-under-the-curve (AUC) statistic for univariate binary outcomes, in the presence of potential verification bias which arises when the true outcome category is unknown. Throughout, we illustrate the proposed methods using a stem cell transplant dataset.

✉ cal373@mail.harvard.edu

INFERENCE ON THE WIN RATIO FOR CLUSTERED SEMI-COMPETING RISK DATA

Di Zhang*, *University of Pittsburgh*

Jong-Hyeon Jeong, *University of Pittsburgh*

The cluster randomization has been increasingly popular for pragmatic clinical trials. The main advantages of using the cluster randomization include minimizing experimental contamination, and increasing the administrative efficiency. Semi-competing risks data arise when a terminal event censors a nonterminal event, but not vice versa. Abundant literature exist on model-based methods to analyze such data. The win ratio is a purely nonparametric summary measure of a group effect in semi-competing risks data accounting for priorities of composite endpoints. In this paper, we propose inference on the win ratio for clustered semi-competing risks data, which can be formulated as the ratio of two clustered U-statistics. First the asymptotic joint distribution of the two clustered U-statistics is derived by using the Cramer-Wold device, their variance and covariance estimators are evaluated, and then a test statistic for the win ratio for clustered semi-competing risks data is constructed. Simulation results are presented to assess type I error probabilities and powers of the test statistic.

✉ diz11@pitt.edu

COMPETING RISKS REGRESSION FOR CASE-COHORT DESIGN

Soyoung Kim*, *Medical College of Wisconsin*

Yayun Xu, *Medical College of Wisconsin*

Mei-Jie Zhang, *Medical College of Wisconsin*

Kwang Woo Ahn, *Medical College of Wisconsin*

The case-cohort study design is an economical means when colleting the expensive covariates in large cohort studies. A case-cohort study design consists of a random sample, called the subcohort as well as all cases or failures. The Fine-Gray proportional hazards model has widely been used for competing risk data to access the effect of covariates on the cumulative incidence function. In this paper, we develop competing risks regression model for case-cohort design and propose more efficient estimators by using extra information for other causes. The proposed estimators are shown to be consistent and asymptotically normally distributed. Simulation studies show that our proposed method performs well and more efficient method using extra information improves efficiency.

✉ skim@mcw.edu

ADJUSTING FOR COVARIATE MEASUREMENT ERROR IN FAILURE TIME ANALYSIS UNDER COMPETING RISKS

Carrie Caswell*, *University of Pennsylvania*

Sharon X. Xie, *University of Pennsylvania*

Time-to-event data in the presence of competing risks has been well studied in recent years. A popular approach to this problem is to model the subdistribution of competing risks with a proportional hazards model, first proposed by Fine and Gray (1999). The estimator resulting from this model does not perform as expected when the covariates are measured with error, which is often the case in biomarker research. We propose a novel method which combines the intuition of Fine and Gray with risk set regression calibration

(Xie, Wang, and Prentice, 2001), which corrects for measurement error in Cox regression by recalibrating at each failure time. We perform simulations to assess under which conditions the Fine and Gray estimator incurs a significant amount of bias in regression coefficients, and demonstrate that our new estimator reduces this bias. We show that the estimator is asymptotically normally distributed and provide a consistent variance estimator. The method is applied to Alzheimer's Disease Neuroimaging Initiative data, which examine the association between measurement error-prone cerebrospinal fluid biomarkers and risk of conversion to Alzheimer's disease.

✉ caswellc@pennmedicine.upenn.edu

JOINT MODELING OF COMPETING RISKS AND CURRENT STATUS DATA: AN APPLICATION TO SPONTANEOUS LABOR STUDY

Youjin Lee*, *Johns Hopkins School of Public Health*

Mei-Cheng Wang, *Johns Hopkins School of Public Health*

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

During the second stage of labor, a cesarean section (CS) or other operational deliveries are encouraged after the guided time set by 'expert consensus'. There may be other benefits from pursuing spontaneous vaginal delivery (SVD) at the cost of allowing more time on labor even beyond the accepted time as CS or other operational deliveries carry their own risks. We compare the risks of SVD and maternal or neonatal morbidities across the duration of second stage labor to find the right time for each individual when these two risks are balanced considering heterogeneity, conditioned on other given baseline covariates. This finding will furnish valuable references for obstetricians about when women should stop pushing. We introduce a semi-parametric joint model which combines competing-risks data for delivery time and current-status data for morbidity with individual-specific frailty, thereby assuring that two different models are independent given observed covariates and individual-level frailty. Our numerical studies which reflect the plausible situations and real data analysis based on more than 18,000 labors will be followed.

✉ ylee160@jhu.edu

64. GENOME-WIDE ASSOCIATION STUDIES

GENETIC ASSOCIATION ANALYSIS OF A MISSING TARGET PHENOTYPE USING MULTIPLE SURROGATE PHENOTYPES

Zachary R. McCaw*, *Harvard School of Public Health*

Xihong Lin, *Harvard School of Public Health*

We consider Genome Wide Association Studies (GWAS) in which the phenotype of primary interest is only ascertained for subjects in a subset of cohorts, while multiple surrogates of the target phenotype are available for subjects in all cohorts. As an example, we consider genetic association analysis of the apnea-hypopnea index (AHI), the gold standard phenotype for diagnosing obstructive sleep apnea. AHI was measured by the Sleep Genetics Epidemiology Consortium (ISGEC), but not in the UK Biobank (UKB), a sample of substantially larger size. Instead, surrogates of AHI, including sleep duration and snoring, are available in UKB. We propose a multivariate association model that jointly considers the surrogate and target phenotypes, and develop inference procedures for the association between genotype and the missing target phenotype. The proposed method accommodates both continuous and binary surrogates, and allows for phenotype-specific regressions. We evaluate the finite sample performance of the proposed methods using simulation studies, and apply the method to genetic association analysis of AHI, and its surrogate sleep phenotypes, using data from the ISGEC and UKB.

✉ zmccaw@g.harvard.edu

ADAPTIVE SNP-SET ASSOCIATION TESTING IN GENERALIZED LINEAR MIXED MODELS WITH APPLICATION TO FAMILY STUDIES

Jun Young Park*, *University of Minnesota*

Chong Wu, *University of Minnesota*

Saonli Basu, *University of Minnesota*

Matt McGue, *University of Minnesota*

Wei Pan, *University of Minnesota*

In genome-wide association studies (GWASs), it has been increasingly recognized that, as a complementary approach to standard single SNP analyses, it may be beneficial to analyze a group of related SNPs together. Among the existent SNP-set association tests, the aSPU test and the aSPUpath test offer a powerful and general approach at the gene- and pathway-levels by data-adaptively combining the results across multiple SNPs (and genes) such that high statistical power can be maintained across a wide range of scenarios. We extend the aSPU and the aSPUpath test to familial data under the framework of the generalized linear mixed models (GLMMs), which can take account of both subject relatedness and possible population structure. Similar to the aSPU test and the aSPUpath test for population-based studies, our methods require only fitting a single GLMM (under the null hypothesis) for all the SNPs, thus are computationally efficient for large GWAS data. We illustrate our approaches in real GWAS data analysis and simulations.

✉ park1131@umn.edu

INCORPORATING GENETIC NETWORKS INTO CASE-CONTROL ASSOCIATION STUDIES WITH HIGH-DIMENSIONAL DNA METHYLATION DATA

Hokeun Sun*, *Pusan National University*

Kipoong Kim, *Pusan National University*

In human genetic association studies with high-dimensional gene expression data, it has been well known that statistical methods utilizing prior biological network like genetic pathways can outperform methods that ignore genetic network structures. In recent epigenetic research on case-control association studies, relatively many statistical methods have been proposed to identify cancer-related CpG sites and the corresponding genes from high-dimensional DNA methylation data. However, most of existing methods are not able to utilize genetic networks. In this article, we propose new approach that combines independent component analysis with network-based regularization to identify outcome-related genes for analysis of high-dimensional DNA methylation data. The proposed approach first captures gene-level signals from multiple CpG sites using independent component analysis and then regularizes them to perform gene selection according to given biological network information. We applied it to the 450K DNA methylation array data of the four breast invasive carcinoma cancer subtypes from the TCGA project.

✉ hsun@pusan.ac.kr

CAUCHY COMBINATION TEST: A POWERFUL TEST WITH ANALYTIC P-VALUE CALCULATION UNDER ARBITRARY DEPENDENCY STRUCTURES

Yaowu Liu*, *Harvard University*

Jun Xie, *Purdue University*

Xihong Lin, *Harvard University*

Combining individual p-values to aggregate multiple small effects has a long-standing interest in statistics, dating back to the classic Fisher's combination test. In modern large-scale data analysis, correlation and sparsity are common features, and efficient computation is a necessary requirement for dealing with massive data. To overcome these challenges, we propose a new test that takes advantage of the Cauchy distribution. We prove a non-asymptotic result that the tail of the null distribution of our proposed test statistic can be well approximated by a Cauchy distribution under arbitrary dependency structures. Based on this theoretical result, the p-value calculation of our proposed test is not only accurate, but also as simple as the classic z-test or t-test, making our test well suited for analyzing massive

data. We further show that the power of the proposed test is asymptotically optimal in a strong sparsity setting.The proposed test has also been applied to a genome-wide association study of Crohn's disease and compared with several existing tests.

✉ yaowuliu615@gmail.com

SIMULTANEOUS SELECTION OF MULTIPLE IMPORTANT SINGLE NUCLEOTIDE POLYMORPHISMS IN FAMILIAL GENOME WIDE ASSOCIATION STUDIES DATA

Subho Majumdar*, *University of Florida*

Saonli Basu, *University of Minnesota*

Snigdhansu Chatterjee, *University of Minnesota*

We propose a resampling-based fast variable selection technique for selecting important Single Nucleotide Polymorphisms (SNP) in multi-marker mixed effect models used in twin studies. To our knowledge, this is the first method of SNP detection in twin studies that uses multi-SNP models. We achieve this through improvements in two aspects. We use the recently proposed e-values framework and a fast and scalable bootstrap procedure to achieve this. We demonstrate the efficacy of our method through simulations and application on a familial GWAS dataset, and detect several SNPs that have potential effect on alcohol consumption in individuals.

✉ smajumdar@ufl.edu

A UNIFIED FRAMEWORK TO PERFORM INFERENCE FOR PLEIOTROPY, MEDIATION, AND REPLICATION IN GENETIC ASSOCIATION STUDIES

Ryan Sun*, *Harvard School of Public Health*

Xihong Lin, *Harvard School of Public Health*

A common challenge in testing for pleiotropy, mediation, and replication in genetic association studies is accounting for a composite null hypothesis. For instance, consider testing for pleiotropic SNPs with two outcomes. The null hypothesis for this problem includes the case where a SNP is associated with no phenotypes as well the cases where a SNP is associated with only one phenotype. A similar situation arises in mediation analysis, where we only want to reject the null hypothesis of no mediation effect when the coefficient of interest is non-zero in both the mediator and outcome models, and in testing for replication, where we want to identify SNPs that demonstrate association across multiple GWAS. Popular approaches - such as the Sobel test or maximum p-value test for mediation - often produce highly conservative inference, resulting in lower power. Borrowing ideas from replicability analysis, we extend an empirical Bayes framework to allow for inference in all three settings. Simulation demonstrates that our approach can control false discovery proportion across various scenarios, and we apply our methods to GWAS of lung cancer and heart disease.

✉ ryanrsun@gmail.com

PENALIZED INFERENCE WITH MANTEL'S TEST FOR MULTI-MODAL ASSOCIATIONS

Dustin S. Pluta* •, *University of California, Irvine*

Tong Shen, *University of California, Irvine*

Hernando Ombao, *King Abdullah University of Science and Technology*

Zhaoxia Yu, *University of California, Irvine*

Mantel's test (MT) for association is conducted by testing the linear relationship of similarity of all pairs of subjects between two observational domains. Motivated by applications to neuroimaging and genetics data, this paper develops a framework based on MT, from which connections between several well known models and MT are established. Inspired by penalization methods for prediction, we propose the use of shrinkage parameters in the calculation of similarity in order to improve the statistical power of MT. Using the concept of variance explained, we provide a heuristic for choosing reasonable tuning parameters for testing with ridge penalized similarity. Through examination of the Mantel test statistics for kernels

related to fixed effects, random effects, and ridge regression models, we unify the score tests of these three models as a single family of tests parameterized by the ridge penalty term. The performance of these tests is compared on simulated data, and illustrated through application to a real neuroimaging and genetics data set.

✉ dpluta@uci.edu

65. META-ANALYSIS

CAUSAL EFFECTS IN META-ANALYSIS OF RANDOMIZED CLINICAL TRIALS WITH NONCOMPLIANCE: A BAYESIAN HIERARCHICAL MODEL

Jincheng Zhou*, *University of Minnesota*

M. Fareed Khan Suri, *University of Minnesota*

Haitao Chu, *University of Minnesota*

Noncompliance to assigned treatments is a common challenge in the analysis and interpretation of randomized clinical trials. The complier average causal effect (CACE) estimation approach provides a useful tool for addressing noncompliance, where CACE is defined as the average difference in potential outcomes for the response in a subpopulation of subjects who comply with their assigned treatments. In this article, we present a Bayesian hierarchical model to esti-mating the CACE in a meta-analysis or a multi-center randomized clinical trial where the com-pliance information may be heterogeneous among studies or centers. Between-study (or center) heterogeneity are taken into account with study-specific random effects. The results are illus-trated through reanalyzing a meta-analysis comparing epidural analgesia to no or other analge-sia in labor on the outcome of cesarean section, where noncompliance rates vary across studies. Finally, we conduct comprehensive simulations to evaluate the performance of the proposed approach, and illustrate the importance of including appropriate random effects and the impact of over- and under- fitting.

✉ jzhou@umn.edu

QUANTIFYING AND PRESENTING OVERALL EVIDENCE IN NETWORK META-ANALYSIS

Lifeng Lin*, *Florida State University*

Network meta-analysis (NMA) has been popular to compare multiple treatments by synthesizing direct and indirect evidence. Many studies did not properly report the evidence of treatment comparisons and show the comparison structure. Also, nearly all treatment networks presented only direct evidence, not overall evidence that reflects the advantage of performing NMAs. We classify treatment networks into three types under different assumptions; they include networks with each edge's width proportional to the corresponding number of studies, sample size, and precision. Three new measures are proposed to quantify overall evidence gained in NMAs. They permit audience to intuitively evaluate the benefit from NMAs. We use some case studies to show their calculation and interpretation. Networks may look very differently when different measures were used to present the evidence. The proposed measures provided clear comparisons between overall evidence of all comparisons. Some comparisons were benefited little from NMAs. Researchers are encouraged to preliminarily present overall evidence of all treatment comparisons, so that audience can evaluate the benefit of performing NMAs.

✉ llin4@fsu.edu

CORRECTING FOR EXPOSURE MISCLASSIFICATION IN META-ANALYSIS: A BAYESIAN APPROACH

Qinshu Lian* •, *University of Minnesota*

James S. Hodges, *University of Minnesota*

Richard Maclehose, *University of Minnesota*

Haitao Chu, *University of Minnesota*

In observational studies, misclassification of exposure measurement is ubiquitous and can substantially bias the association between an outcome and an exposure. Although misclassification in a single observational study has been well studied, few papers considered it in a meta-analysis.

Meta-analyses of observational studies provide important evidence for health policy decisions, especially when large randomized controlled trials are unavailable. It is imperative to account properly for misclassification in a meta-analysis to obtain valid point and interval estimates. In this paper, we propose a novel Bayesian approach to filling this methodological gap. We simultaneously synthesize two meta-analyses, with one on the association between a misclassified exposure and an outcome (main studies), and the other on the association between the misclassified exposure and the true exposure (validation studies). We extend the current scope of using external validation data by relaxing the transportability assumption by means of random effects models. Our model accounts for heterogeneity between studies and allows different studies to have different exposure measurements.

✉ lianx025@umn.edu

EAMA: EMPIRICALLY ADJUSTED META-ANALYSIS FOR LARGE-SCALE SIMULTANEOUS HYPOTHESIS TESTING IN GENOMIC EXPERIMENTS

Sinjini Sikdar*, *National Institute of Environmental Health Sciences, National Institutes of Health*

Somnath Datta, *University of Florida*

Susmita Datta, *University of Florida*

Recent developments in high throughput genomic assays have opened up the possibility of testing hundreds and thousands of genes simultaneously. However, adhering to the regular statistical assumptions regarding the null distributions of test statistics in such large-scale multiple testing frameworks has the potential of leading to incorrect significance testing results and biased inference. This problem gets even worse when one combines results from different independent genomic experiments with a possibility of ending up with gross false discoveries of significant genes. In this project, we develop a novel and very useful meta-analysis method of combining p-values from different independent experiments involving large-scale multiple testing frameworks, through empirical adjustments of the individual test statistics and p-values. Through multiple simulation studies and real genomic datasets we show that our method outperforms the standard meta-analysis approach of significance testing in terms of accurately identifying the truly significant set of genes, especially in presence of hidden confounding covariates.

✉ sinjini.sikdar@nih.gov

META-ANALYSIS OF INCIDENCE OF RARE EVENTS USING INDIVIDUAL PATIENT-LEVEL DATA

Yan Ma*, *The George Washington University*

Chen Chen, *The George Washington University*

Yong Ma, *U.S. Food and Drug Administration*

Individual participant or patient data (IPD) meta-analysis (M-A) is an increasingly popular approach, which provides individual data rather than summary statistics compared to a study-level M-A. By pooling data across multiple studies, meta-analysis increases statistical power. However, existing IPD M-A methods make inferences based on large sample theory and have been criticized for generating biased results when handling rare events/outcomes, such as adverse events in drug safety studies. We propose an exact statistical method based on a Poisson-Gamma hierarchical model in a Bayesian framework to take rare events into account. In addition to the development of the theoretical methodology, we also conduct a simulation study to examine and compare the proposed method with other approaches: the naïve approach of simply combining data from all available studies ignoring the between-study heterogeneity, and a random effects model built on large number theory.

✉ yanma@gwu.edu

MULTILEVEL MIXED-EFFECT STATISTICAL MODELS FOR INDIVIDUAL PARTICIPANT DATA META-ANALYSIS

Ying Zhang*, *The Pennsylvania State Health Milton S. Hershey Medical Center*

Vernon M. Chinchilli, *The Pennsylvania State Health Milton S. Hershey Medical Center*

Individual participant data (IPD) meta-analysis that combines and analyzes raw data from studies has been suggested to be more powerful and flexible compared with meta-analysis based on summary statistics. We propose a statistical model that is a combination of generalized linear mixed-effect models and multilevel models such that the new models contain (a) fixed and random effects for the longitudinal data from each participant within a study, and (b) fixed and random effects for a study. The models can accommodate outcome variables that are continuous or from an exponential family. We derive the estimators for fixed-effect parameters and variance-covariance parameters. To evaluate the proposed models, we performed a simulation study in which we generated multicenter longitudinal clinical data to mimic clinical studies investigating a treatment effect and then applied the proposed models, 3-level and 4-level mixed-effect models. Compared with naïve models, the proposed models generally improved the precision, as indicated by smaller estimates of standard deviations of fixed-effect parameters, and provided more accurate estimates of variance-covariance parameters.

✉ ymz5137@psu.edu

TESTING EQUALITY OF MEANS IN PARTIALLY PAIRED DATA WITH INCOMPLETENESS IN SINGLE RESPONSE

Qianya Qi*, *State University of New York at Buffalo*

Li Yan, *Roswell Park Cancer Institute*

Lili Tian, *State University of New York at Buffalo*

In testing differentially expressed genes between tumor and healthy tissues, data are usually collected in paired form. However, incomplete paired data often occur. While extensive statistical researches exist for paired data with incompleteness in both arms, hardly any recent work can be found on paired data with incompleteness in single arm. In this talk, we present some methods for testing hypothesis for such data. Simulation studies demonstrate that the proposed methods can maintain type I error well and have good power property. A real data set from The Cancer Genome Atlas (TCGA) breast cancer study is analyzed using the proposed methods. The proposed methods should have wide applicability in practical fields.

✉ qianyaqi@buffalo.edu

66. MISSING DATA METHODS

COARSENED PROPENSITY SCORES AND HYBRID ESTIMATORS FOR MISSING DATA AND CAUSAL INFERENCE

Jie Zhou*, *U.S. Food and Drug Administration*

Zhiwei Zhang, *University of California, Riverside*

Zhaohai Li, *The George Washington University*

Jun Zhang, *Shanghai Jiaotong University School of Medicine*

In the areas of missing data and causal inference, there is great interest in doubly robust (DR) estimators that involve both an outcome regression (OR) model and a propensity score (PS) model. These DR estimators are consistent and asymptotically normal if either model is correctly specified. Despite their theoretical appeal, the practical utility of DR estimators has been disputed. One of the major concerns is the possibility of erratic estimates resulting from near zero denominators. In contrast, the usual OR estimator is efficient when the OR model is correct and generally more stable, although it can be biased when the OR model is incorrect. In light of the unique advantages of the OR and DR estimators, we propose a class of hybrid estimators that attempt to strike a reasonable balance between the OR and DR estimators. These hybrid estimators are based on coarsened PS estimates, which are less likely to take extreme values and less sensitive to misspecification of the PS model. The proposed estimators are compared to existing estimators in simulation studies and illustrated with real data from a large observational study.

✉ jack.zhou@fda.hhs.gov

EMPIRICAL-LIKELIHOOD-BASED CRITERIA FOR JOINT MODEL SELECTION ON WEIGHTED GENERALIZED ESTIMATING EQUATION ANALYSIS OF LONGITUDINAL DATA WITH DROPOUT MISSINGNESS

Chixiang Chen*, *The Pennsylvania State University*

Ming Wang, *The Pennsylvania State University*

Longitudinal data are common in clinical trials or observational studies, and the outcomes with missing data due to dropouts are always encountered. Weighted generalized estimating equations (WGEE) was proposed for such context under the assumption of missing at random. Of note is that correctly specifying marginal mean regression and correlation structure can lead to the most efficient estimators for statistical inference; however, there exist limited work developing joint model selection criteria and the existing criteria have restrictions with unsatisfactory performance. In this work, we heuristically propose two innovative criteria, named joint empirical AIC and joint empirical BIC, which jointly select marginal mean model and correlation structure. These empirical-likelihood-based criteria exhibit robustness, flexibility, and outperformance through extensive simulation studies compared to the existing criteria such as weighted quasi-likelihood information criterion (QICW), missing longitudinal information criterion (MLIC). Theoretically its asymptotic behavior and the extension to other estimation equations are also discussed. Lastly, a real data example is presented.

✉ chencxxy@psu.edu

MULTIPLY ROBUST ESTIMATION IN NONPARAMETRIC REGRESSION WITH MISSING DATA

Yilun Sun* •, *University of Michigan*

Lu Wang, *University of Michigan*

Peisong Han, *University of Waterloo*

Nonparametric regression has gained considerable attention in many biomedical studies because of its great flexibility to allow data-driven dependence. To deal with ubiquitous missing data problem, doubly robust estimator has been proposed for nonparametric regression. However, it only allows one model for the missingness mechanism and one model for the outcome regression. We propose multiply robust kernel estimating equations (MRKEEs) for nonparametric regression that can accommodate multiple postulated working models for either the missingness mechanism or the outcome regression, or both. The resulting estimator is consistent if any one of those models is correctly specified. When including correctly specified models for both the missingness mechanism and the outcome regression, the proposed estimator achieves the optimal efficiency within the class of AIPW kernel estimators. We perform simulations to evaluate the finite sample performance of the proposed method and apply it to analyze the data collected on 2078 high-risk cardiac patients enrolled into a cardiac rehabilitation program at the University of Michigan.

✉ yilunsun@umich.edu

CORRECTING BIAS FROM ESTIMATING RISK OF ALZHEIMER'S DISEASE FROM INFORMATIVE CENSORING USING AUXILIARY INFORMATION

Cuiling Wang*, *Albert Einstein College of Medicine*

Charles Hall, *Albert Einstein College of Medicine*

Richard Lipton, *Albert Einstein College of Medicine*

Joe Verghese, *Albert Einstein College of Medicine*

Mindy Katz, *Albert Einstein College of Medicine*

Qi Gao, *Albert Einstein College of Medicine*

Evaluating the risk Alzheimer's disease (AD) and possible risk factors is an important goal in many longitudinal aging studies as AD is a global public health problem of enormous significance. An important challenge facing

these studies is non-random or informative censoring. Participants with poorer health may be more likely to drop out, therefore violates the random censoring assumption which is the basis of regular analyses, which can result in biased results and potential misleading scientific conclusions. Auxiliary data, measures that are associated with the outcome and missing data, allow us to evaluate the random censoring assumption and to eliminate or reduce bias from non-random censored data. We evaluate factors associated with the impact of utilizing auxiliary information through extensive simulation studies, and examine empirically how using longitudinal cognitive data as auxiliary variables may help correct bias from non-random censoring in the estimation of AD risk. The method is applied to data from Einstein Aging Study (EAS).

✉ cuiling.wang@einstein.yu.edu

VARIABLE SELECTION FOR NON-NORMALLY DISTRIBUTED DATA UNDER AN ARBITRARY MISSINGNESS

Yang Yang*, *State University of New York at Buffalo*

Jiwei Zhao, *State University of New York at Buffalo*

Regularized likelihood has been proved to be effective in variable selection. This method has been well developed theoretically and computationally in the past two decades. However, two major problems still exist in practice. One is that the normality of response variable is rare in practice. Clinical data, especially patient reported outcome (PRO), is usually distributed asymmetrically and sometimes finitely, preventing direct application of penalized likelihood. The other one is caused by non-ignorable missing data. The sensitivity of missing mechanism assumption sets obstacles to select variables of interest. To overcome above problems, we first introduce a pseudo likelihood based on proportional likelihood ratio model and then integrate a flexible missing data mechanism. For variable selection, L1 and non-convex penalties will be explored. Cross validation, Bayesian information criterion, and three other stability based techniques are checked for tuning parameter selection. A comprehensive simulation is presented to assess performance of our proposed method. Patient reported pain score from a chondral lesions study is used as response variable in real data study.

✉ yyang39@buffalo.edu

REGRESSION OF OBSERVATIONS BELOW THE LIMIT OF DETECTION: A PSEUDO-VALUE APPROACH

Sandipan Dutta*, *Duke University*

Susan Halabi, *Duke University*

Biomarkers are known to be important in the progression of several diseases, such as cancer. One of the most critical performance metrics for any assay is related to the minimum amount of values that can be detected. Such observations are known as below the limit of detection (LOD) and may have a huge impact on the analysis and interpretation of the data. Deleting these observations may cause loss of information, while retaining them at the LOD can make the data heavily skewed leading to wrong inference. A common approach is to use parametric censored regression model, such as the Tobit regression model, for regressing outcomes below the LOD. Such parametric models, however, heavily depends on the distributional assumptions, and can result in loss of precision in estimating biomarker relationship with the outcome. Instead, we utilize a pseudo-value based regression approach without making any distributional assumptions. We show through simulations that the pseudo-value based regression is more precise and outperforms the parametric models in estimating the biomarker-outcome relationship. We demonstrate the utility of the pseudo-value approach using a real life example.

✉ sandipan.dutta@duke.edu

67. WEARABLE AND PORTABLE DEVICES

ANALYSIS OF TENSOR CUMULANTS AND ITS APPLICATION TO NHANES

Junrui Di*, *Johns Hopkins Bloomberg School of Public Health*

Vadim Zipunnikov, *Johns Hopkins Bloomberg School of Public Health*

Modern technology has generated high dimensional multivariate data in various biomedical applications. One example is continuous monitoring of physical activity (PA) with wearable accelerometers. To address high-dimensionality, dimension reduction techniques such as principal component analysis (PCA) are often applied to explore and analyze these data. Finding components based on the covariance matrix is only adequate to characterize multivariate Gaussian distribution. However, accelerometry data often exhibits significant deviation from Gaussian distribution with high skewness and kurtosis. To address it, we propose Analysis of Tensor Cumulants (ATC). It constructs 3rd and 4th order cumulant tensors to capture higher order information. The cumulant tensors are then decomposed via symmetric tensor decompositions. The proposed approach extends PCA by conducting decomposition of the observed data on the original scale and by accounting for the non-Gaussianity. We apply ATC to accelerometry data of 3400 participants of 2003-2006 National Health and Nutrition Examination Survey and explore associations between ATC estimated diurnal patterns of PA and the follow-up mortality.

✉ jdi2@jhu.edu

UNSUPERVISED CLUSTERING OF PHYSICAL ACTIVITIES AND ITS APPLICATION IN HEALTH STUDIES

Jiawei Bai*, *Johns Hopkins University*

Ciprian M. Crainiceanu, *Johns Hopkins University*

The time spent in different physical activities per day was found to be highly associated with many health factors, and thus, accurately measuring the time is critical. Currently many supervised learning methods provided high prediction accuracy for activity type, but their usage were limited to several key known activities, such as sitting still and walking. Many less common or not-well-defined activities were ignored due to the difficulty of establishing reliable training data. We proposed an unsupervised learning method to extract a set dominating patterns of signal from the acceleration time series. We further investigated the interpretation of these patterns and established a relationship between them and some well-defined activities. Using this method, we avoided manually defining types or categories of activity and were still able to investigate the association between the time spent in each category and health factors.

✉ jbai@jhsph.edu

PENALIZED AUGMENTED ESTIMATING EQUATIONS FOR MODELING WEARABLE SENSOR DATA WITH INFORMATIVE OBSERVATION TIMES AND CENSORING TIME

Jaejoon Song*, *University of Texas MD Anderson Cancer Center*

Michael D. Swartz, *University of Texas Health Science Center at Houston*

José-Miguel Yamal, *University of Texas Health Science Center at Houston*

Kelley Pettee Gabriel, *University of Texas Health Science Center at Houston*

Karen Basen-Engquist, *University of Texas MD Anderson Cancer Center*

Large population-based studies such as the National Health and Nutrition Examination Survey (NHANES) have included wearable sensors to collect longitudinal estimates of physical activity data in free-living settings. However, the quality of data collected using such wearable sensors often rely on participant fidelity. In other words, there may be high variability in the device wear times during waking hours within- and between- individuals over the scheduled measurement days. In addition, since device wear relies on participants' free will, wear times may be associated with the measurement outcomes (informative observation times and informative censoring). We propose a penalized semiparametric model to explore potential correlates to the wearable sensor measured outcome, while accounting for the missing data features from these data. In a simulation study, our proposed method was unbiased under informative observation times and informative censoring, and showed high accuracy in correctly selecting the true predictors in the model. Our method was applied to real data from the NHANES 2003-04, to explore factors associated with real world physical activity.

✉ jaejoonsong@gmail.com

CHANGE POINT DETECTION FOR MULTIVARIATE DIGITAL PHENOTYPES

Ian J. Barnett*, *University of Pennsylvania*

Traits related to mobility, sociability, sleep, and other aspects of human behavior can be quantified based on smartphone sensor data from every day use. Monitoring these traits can inform interventions in patient populations with suicidal ideation, substance use disorders, and other psychiatric disorders. This data can be represented as a multivariate time series, where change point detection methods can be used to prompt interventions. New methods are needed capable of accounting for both the complex distributions of these behavioral traits as well as high amounts of missing data. We propose a doubly robust nonparametric data transformation as well as a variance component test for change point detection in this setting. The power of this approach is demonstrated relative to competing methods through simulation as well as to predict relapse in a cohort of patients with schizophrenia.

✉ ibarnett@pennmedicine.upenn.edu

AUTOMATED LONGITUDINAL LATENT INTERVAL ESTIMATION WITH APPLICATIONS TO SLEEP

Patrick Staples*, *Harvard School of Public Health*

Estimating sleep over time is difficult due to the paucity of unobtrusive longitudinal data related to sleep. We propose an approach using digital phenotyping, or the moment-by-moment quantification of individual-level phenotype in-situ using personal digital devices, in particular smartphones. Although smartphone ownership and usage continues to increase, accounting for the indirect relationship between smartphone activity and sleep status presents unique challenges, for which strong but potentially testable assumptions must be made. In this presentation, we introduce an unsupervised, subject-specific, longitudinal, likelihood-based framework for estimating the latent daily onset of sleep and waking from arbitrary smartphone activity data and longitudinal covariates. We compare the empirical and theoretical bias and variance of parameter estimates via simulation. We apply our method to a cohort of healthy students at Harvard College, and estimate the method's accuracy against sleep estimates derived from FDA-approved actigraphy devices. We also apply the method to several studies using Beiwe, our digital phenotyping platform, in a range of clinical contexts.

✉ patrickstaples@fas.harvard.edu

68. CHALLENGES, OPPORTUNITIES, AND METHODS FOR LEARNING FROM LARGE-SCALE ELECTRONIC HEALTH RECORDS DATABASES

ADJUSTING FOR SELECTION BIAS IN ELECTRONIC HEALTH RECORDS-BASED RESEARCH

Sebastien Haneuse*, *Harvard School of Public Health*

Sarah Peskoe, *Duke University*

David Arterburn, *Kaiser Permanente Washington Health Research Institute*

Michael Daniels, *University of Florida*

While EHR data provide unique opportunities for public health research, selection due to incomplete data is an underappreciated source of bias. When framed as a missing-data problem, standard methods could be applied, although these typically fail to acknowledge the often-complex interplay of clinical decisions made by patients, providers, and the health system, required for data to be complete. As such, residual selection bias may remain. Building on a recently-proposed framework for characterizing how data arise in EHR-based studies, we develop and evaluate a statistical framework for regression modeling based on inverse probability weighting that adjusts for selection bias in the complex setting of EHR-based research. We show that the resulting estimator is consistent and asymptotically Normal, and derive the form of the asymptotic variance. We use simulations to highlight the potential for bias when standard approaches are used to account for selection bias, and evaluate the small-sample operating characteristics of the proposed framework. Finally, the methods are illustrated using data from an on-going, multi-site EHR-based study of bariatric surgery on BMI.

✉ shaneuse@hsph.harvard.edu

ACCOUNTING FOR INFORMATIVE PRESENCE BIAS AND LACK OF PORTABILITY IN EHR-DERIVED PHENOTYPES

Rebecca A. Hubbard*, *University of Pennsylvania*

Joanna Horton, *University of Pennsylvania*

Jing Huang, *University of Pennsylvania*

Yong Chen, *University of Pennsylvania*

Electronic Health Records (EHR) include a wide variety of clinical data that can be used to describe patient phenotypes. As a result, phenotyping algorithms have been developed for many conditions of interest. Despite their wide availability, phenotypes developed in one EHR often perform poorly in others, a phenomenon referred to as lack of portability. EHR-based phenotyping algorithms also must address bias due to "informative presence," in which data elements are less likely to be missing for individuals with the condition of interest because they interact with the healthcare system more frequently than unaffected individuals. To address these issues, we propose a hierarchical Bayesian model that incorporates information from existing phenotyping algorithms via prior distributions and allows for healthcare system-specific variation in algorithm performance. Motivated by a multi-site study of pediatric type 2 diabetes (T2D), we conducted simulation studies to evaluate the proposed approach. We applied new and alternative approaches to evaluate the prevalence of T2D and associations between T2D and adverse health outcomes.

✉ rhubb@pennmedicine.upenn.edu

LEARNING INDIVIDUALIZED TREATMENT RULES FROM ELECTRONIC HEALTH RECORDS

Yuanjia Wang*, *Columbia University*

Current guidelines for treatment decision making largely rely on data from randomized controlled trials (RCT) studying average treatment effects. They may be inadequate to make individualized treatment decisions in real-world settings. Large-scale electronic health records (EHR) data provide unprecedented opportunities to learn individual-

ized treatment rule (ITR) depending on patient-specific characteristics from real world data. We propose a machine learning approach to estimate ITRs, referred to as the matched learning (M-Learning). This new learning method performs matching instead of inverse probability weighting to more accurately estimate an individual's treatment response under alternative treatments and alleviate confounding in observational studies. A matching function is proposed to compare outcomes for matched pairs where various types of outcomes (including continuous, ordinal and discrete responses) can easily be accommodated under a unified framework. We conduct extensive simulation studies and apply our method to a study of optimal second-line treatments for type 2 diabetes patients using EHRs.

✉ yw2016@cumc.columbia.edu

USING ELECTRONIC HEALTH RECORDS DATA TO TARGET SUICIDE PREVENTION CARE

Susan M. Shortreed*, *Kaiser Permanente Washington Health Research Institute*

Gregory E. Simon, *Kaiser Permanente Washington Health Research Institute*

Eric Johnson, *Kaiser Permanente Washington Health Research Institute*

Jean M. Lawrence, *Kaiser Permanente Southern California*

Rebecca C. Rossum, *HealthPartners Institute*

Brian Ahmedani, *Henry Ford Health System*

Frances M. Lynch, *Kaiser Permanente Northwest Center for Health Research*

Arne Beck, *Kaiser Permanente Colorado Institute for Health Research*

Rebecca Ziebell, *Kaiser Permanente Washington Health Research Institute*

Robert B. Penfold, *Kaiser Permanente Washington Health Research Institute*

Suicide is the 10th leading cause of death in the US. Effective suicide prevention interventions exist, but are often resource intensive. Successful identification of those at increased risk for suicide attempt and death makes implementing suicide prevention interventions on a large scale feasible. Electronic health records (EHRs) contain vast amounts of information on the health care patients have sought and received in real medical settings. We will present work that uses EHR data to identify individuals at risk of suicide. We will discuss the different types of EHR data that may be available to different systems (e.g. administrative/claims data versus patient reported outcomes) and how this data can be complementary. We will highlight the statistical and computational challenges we faced conducting scientific research using EHR data on millions of patients. We will illustrate the potential for using EHR data to advance medicine with the results of this research project that used data gathered from clinical visits and administrative data to identify individuals at increased risk of suicide in order to better target care.

✉ shortreed.s@ghc.org

69. GEOMETRY AND TOPOLOGY IN STATISTICAL INFERENCE

MANIFOLD LEARNING ON FIBRE BUNDLES

Tingran Gao*, *University of Chicago*

Jacek Brodzki, *University of Southampton*

Sayan Mukherjee, *Duke University*

We develop a geometric framework, based on the classical theory of fibre bundles, to characterize the cohomological nature of a large class of synchronization-type problems in the context of graph inference and combinatorial optimization. In this type of problems, the pairwise interaction between adjacent vertices in the graph is of a "non-scalar" nature, typically taking values in a group or groupoid; the "consistency" among these non-scalar pairwise interactions provide information for the dataset from which

the graph is constructed. We model these data as a fibre bundle equipped with a connection, and consider a horizontal diffusion process on the fibre bundle driven by a standard diffusion process on the base manifold of the fibre bundle; the spectral information of the horizontal diffusion decouples the base manifold structure from the observed non-scalar pairwise interactions. We demonstrate an application of this framework on evolutionary anthropology.

✉ tingrangao@galton.uchicago.edu

HYPOTHESIS TESTING FOR SPATIALLY COMPLEX DATA USING PERSISTENT HOMOLOGY SUMMARIES

Jessi Cisewski-Kehe*, *Yale University*

Data exhibiting complicated spatial structures are common in many areas of science (e.g. cosmology, biology), but can be difficult to analyze. Persistent homology offers a new way to represent, visualize, and interpret complex data by extracting topological features, which can be used to infer properties of the underlying structures. Persistent homology can be thought of as finding different ordered holes in data where dimension 0 holes are connected components, dimension 1 holes are loops, dimension 2 holes are voids, and so on. The summary diagram is called a "persistence diagram" -- a barcode plot conveys the same information in a different way. These topological summaries can be used as inputs in inference tasks (e.g. hypothesis tests). The randomness in the data due to measurement error or topological noise is transferred to randomness in these topological summaries, which provides an infrastructure for inference. This allows for statistical comparisons between spatially complex datasets. We present several possible test statistics for two-sample hypothesis tests using persistence diagrams.

✉ jessica.cisewski@yale.edu

FUNCTIONAL DATA ANALYSIS USING A TOPOLOGICAL SUMMARY STATISTIC: THE SMOOTH EULER CHARACTERISTIC TRANSFORM

Lorin Crawford*, *Brown University School of Public Health*

Anthea Monod, *Columbia University*

Andrew X. Chen, *Columbia University*

Sayan Mukherjee, *Duke University*

Raúl Rabadán, *Columbia University*

We introduce a novel statistic, the smooth Euler characteristic transform (SECT), which is designed to integrate shape information into regression models by representing shapes and surfaces as a collection of curves. Due to its well-defined inner product structure, the SECT can be used in a wider range of functional and nonparametric modeling approaches than other previously proposed topological summary statistics. We illustrate the utility of the SECT in a radiomics context by showing that the topological quantification of tumors, assayed by magnetic resonance imaging (MRI), are better predictors of clinical outcomes in patients with glioblastoma multiforme (GBM). We show that SECT features alone explain more of the variance in patient survival than gene expression, volumetric features, and morphometric features.

✉ lorin_crawford@brown.edu

GEOMETRIC METHODS FOR MODELING TIME EVOLUTION IN HUMAN MICROBIOTA

Justin D. Silverman*, *Duke University*

Sayan Mukherjee, *Duke University*

Lawrence A. David, *Duke University*

Within the biomedical community there is an increasing recognition of the importance that host-associated microbes play in both human health and disease. Moreover, there has been much excitement over the insights that can be

obtained from longitudinal measurements of these microbial communities; however, due to statistical limitations appropriate models have been lacking. Host microbiota are typically measured using high-throughput DNA sequencing which results in counts for different species. Relative abundances are then estimated from these counts. In addition, due to technological limitations the total number of counts per sample is often small compared to the distribution of species relative abundances leading to datasets with many zero or small counts. With such data, models that incorporate the sampling variability are essential. To accommodate time-series modeling of host microbiota, a multinomial-normal-on-the-simplex generalized dynamic linear model has been developed. Using a combination of both real and simulated datasets we demonstrate that this modeling framework enables accurate inference of the effects of prebiotic treatments in microbiota time-series.

✉ Justin.Silverman@duke.edu

70. PREDICTIVE MODELING OF ACCELEROMETRY, ELECTRONIC DIARIES, AND PASSIVELY RECORDED VOICE DATA

HOW RICH IS THE RAW ACCELEROMETRY DATA? WALKING VS. STAIR CLIMBING

Jaroslaw Harezlak*, *Indiana University Fairbanks School of Public Health*

William Fadel, *Indiana University Fairbanks School of Public Health*

Jacek Urbanek, *Johns Hopkins University School of Medicine*

Xiaochun Li, *Indiana University School of Medicine*

Steven Albertson, *Indiana University Purdue University, Indianapolis*

Wearable accelerometers offer a noninvasive measure of physical activity (PA). They record unlabeled high frequency three-dimensional time series data. Among many human activities, walking is the most common moderate level PA. Our work addresses the classification of walking into level walking, descending stairs and ascending stairs. We apply our method based on the extracted short-time interpretable features arising from the Fourier and wavelet transforms to data collected on N=32 middle-aged participants. We build subject-specific and group-level classification models utilizing a tree-based classifier. We evaluate the effects of sensor location and tuning parameters on the classification accuracy of these models. In the group-level classification setting, we propose a robust feature normalization approach and evaluate its performance. In summary, our work provides a framework for better feature extraction and use of the raw accelerometry data to differentiate among different walking modalities. We show that both at a subject-specific level and at a group level, overall classification accuracy is above 80% indicating excellent performance of our method.

✉ harezlak@iu.edu

WEEK-TO-WEEK ACTIGRAPHY TRACKING OF CLINICAL POPULATIONS USING MULTI-DOMAIN DECOMPOSITION

Vadim Zipunnikov*, *Johns Hopkins Bloomberg School of Public Health*

To track health status of patients during pre- and post-intervention periods (such as surgery, organ replacement, hospitalization, etc.) many ongoing clinical trials and studies ask subjects to wear sleep or activity trackers over many weeks and months. Thus, it is important to estimate within-subject trajectories of health status derived from wearables. We propose a novel real-time actigraphy-based score that combines multiple features of three domains including physical activity, sleep, and circadian rhythmicity and captures week-to-week trajectories of subject-status defined by these domains. The approach takes raw actigraphy data, extracts multiple features for each domain and each week, aggregates these features into a score that characterizes within-subject week-to-week changes in subjects status. We demonstrate that the proposed score is highly

efficient both for tracking subject' status after the adverse events as well as for prediction of events in the study of 54 individuals diagnosed with congestive heart failure who wore Actical tracking device over 6 to 10 month periods.

✉ vadim.zipunnikov@gmail.com

PREDICTING MOOD STATES IN BIPOLAR DISORDER FROM ANALYSES OF ACOUSTIC PATTERNS RECORDED FROM MOBILE TELEPHONE CALLS

Melvin Mcinnis*, *University of Michigan*

Soheil Khorram, *University of Michigan*

John Gideon, *University of Michigan*

Emily Mower Provost, *University of Michigan*

Bipolar disorder (BP) is a psychiatric disease characterized by pathological mood swings, ranging from mania to depression. PRIORI analyzes the acoustic components of speech collected from smartphones to predict mood states for individuals with BP. Ground truth mood labels were assessed in weekly assessment calls with standardized clinical assessment instruments (HamD and YMRS). Our pilot investigation shows that a SVM-based classifier trained on rhythm features can recognize manic and depressive mood states with the AUCs of 0.57+/-0.25 and 0.64+/-0.14, respectively. Because of variability in recording, a preprocessing pipeline consisting of three modules: RBAR declipping, combo-SAD segmentation, and speaker normalization. This approach significantly increases the mania and depression recognition AUCs to 0.72+/-0.20 and 075+/-0.14 , respectively. Our experiments show that the fusion of subject-dependent and population general systems significantly outperforms both single systems.

✉ mmcinnis@med.umich.edu

IMPROVED MODELING OF SMARTPHONE-BASED ECOLOGICAL MOMENTARY ASSESSMENT DATA FOR DIETARY LAPSE PREDICTION

Fengqing Zhang*, *Drexel University*

Tinashe M. Tapera, *Drexel University*

Stephanie P. Goldstein, *Drexel University*

Evan M. Forman, *Drexel University*

Obesity, a condition present in 35% of US adults, increases risk for numerous diseases, and reduces quality of life. Participants in weight loss programs struggle to remain adherent to a dietary prescription. Specific moments of inadherence, known as dietary lapses, are the cause of weight control failure. We developed a smartphone app that utilizes just-in-time adaptive intervention and machine learning to predict and prevent dietary lapses. Users were repeatedly prompted to enter information about lapses and a set of potentially triggering factors (e.g., mood) using a repeated sampling method called ecological momentary assessment. The resulting data have an unbalanced ratio of lapses to non-lapses approximately 1:12. Classification of data with imbalanced class distribution is challenging. To this end, we developed a cost-sensitive ensemble model as a meta-technique to combine multiple weak classifiers and introduce cost items into the learning framework. We also designed a neighborhood based balancing strategy to redefine the training set for a given test set. Results showed that the proposed model works efficiently and effectively for lapse prediction.

✉ fengqingzoezhang@gmail.com

7I. INTEGRATIVE ANALYSIS FOR BRAIN IMAGING STUDIES

INTERMODAL COUPLING ANALYTICS FOR MULTIMODAL NEUROIMAGING STUDIES

Russell T. Shinohara*, *University of Pennsylvania*

A proliferation of MRI-based neuroimaging modalities now allows measurement of diverse features of brain structure, function, and connectivity during the critical period of adolescent brain development. However, the vast majority of developmental imaging studies use data from each neuroimaging modality independently. As such, most developmental studies have not considered, or have been unable to consider, potentially rich information regarding relationships between imaging phenotypes. At present, it remains unknown how local patterns of structure and function are related, how this relationship changes through adolescence as part of brain development, and how developmental pathology may impact such relationships. Here, we propose to measure the relationships between measures of brain structure, function, and connectivity during adolescent brain development by developing novel, robust analytic tools for describing relationships among imaging phenotypes. Our over-arching hypothesis is that such relationships between imaging features will provide uniquely informative data regarding brain health, over and above the content of data from each modality when considered in isolation.

✉ rshi@upenn.edu

JOINT ANALYSIS OF MULTIMODAL IMAGING DATA VIA NESTED COPULAS

Jian Kang*, *University of Michigan*

Peter X.K. Song, *University of Michigan*

In this work, we develop a modeling framework for joint analysis of multimodal imaging data via nested copulas, which appropriately characterize the dependence among multiple imaging modalities as well as the complex spatial correlations across different locations. We have three levels of hierarchy. At level 1, we specify the marginal distribution of the modality specific imaging outcome at each location (e.g. voxel or region of interest) in the brain. At level 2, at each location of the brain, we model the joint distribution of the multimodal imaging data via modality-dependent copulas. At level 3, we resort to another location-dependent copula construct the joint distribution of multimodal imaging outcomes over space. The modality-dependent copulas are nested in the location-dependent copula. We study the theoretical properties of the proposed method and develop efficient model inference algorithms from both Bayesian and frequentist perspectives. We illustrate the performance of the proposed method via simulation studies and a joint analysis of resting-state functional magnetic resonance imaging (fMRI) data and diffusion tensor imaging (DTI) data.

✉ jiankang@umich.edu

A BAYESIAN PREDICTIVE MODEL FOR IMAGING GENETICS WITH APPLICATION TO SCHIZOPHRENIA

Francesco C. Stingo*, *University of Florence*

Thierry Chekouo, *University of Minnesota*

Michele Guindani, *University of California, Irvine*

Kim-Anh Do, *University of Texas MD Anderson Cancer Center*

Imaging genetics has rapidly emerged as a promising approach for investigating the genetic determinants of brain mechanisms that underlie an individual's behavior or psychiatric condition. In particular, for early detection and targeted treatment of schizophrenia, it is of high clinical relevance to identify genetic variants and imaging-based biomarkers that can be used as diagnostic markers, in addition to commonly used symptom-based assessments. By combining single-nucleotide polymorphism (SNP) arrays and functional magnetic resonance imaging (fMRI), we propose an integrative Bayesian risk prediction model that allows us to discriminate between individuals with

schizophrenia and healthy controls, based on a sparse set of discriminatory regions of interest (ROIs) and SNPs. Inference on a regulatory network between SNPs and ROI intensities (ROI–SNP network) is used in a single modeling framework to inform the selection of the discriminatory ROIs and SNPs. We found our approach to outperform competing methods that do not link the ROI–SNP network to the selection of discriminatory markers.

✉ fra.stingo@gmail.com

INTEGRATIVE METHODS FOR FUNCTIONAL AND STRUCTURAL CONNECTIVITY

DuBois Bowman*, *Columbia University*

There is emerging promise in combining data from different imaging modalities to determine connectivity in the human brain and its role in various disease processes. There are numerous challenges with such integrated approaches, including specification of flexible and tenable modeling assumptions, correspondence of functional and structural linkages, and the potentially massive number of pairwise associations, to name a few. In this talk, I will present some useful approaches that target combining functional and structural data to assess functional connectivity and to determine brain features that reveal a neurological disease process, namely Parkinson's disease. The proposed methods are relatively straightforward to implement and have revealed good performance in simulation studies and in applications to various neuroimaging data sets.

✉ dubois.bowman@columbia.edu

72. NOVEL EXTENSIONS AND APPLICATIONS OF CAUSAL INFERENCE MODELS

MODEL-BASED STANDARDIZATION USING AN OUTCOME MODEL WITH RANDOM EFFECTS

Babette Anne Brumback*, *University of Florida*

Zhongkai Wang, *University of Florida*

Adel Alrwisan, *University of Florida*

Almut Winterstein, *University of Florida*

Model-based standardization uses a statistical model to estimate a standardized, or unconfounded, population-average effect. With it, one can compare groups had the distribution of confounders been identical in both groups to that of the standard population. Typically, model-based standardization relies on either an exposure model or an outcome model. Inverse-probability of treatment weighted estimation of marginal structural model parameters can be viewed as model-based standardization with an exposure model. We develop an approach based on an outcome model, in which the mean of the outcome is modeled conditional on the exposure and the confounders. In our approach, there is a confounder that clusters the observations into a large number of categories, e.g., zip code or individual. We treat the parameters for the clusters as random effects. We use a between-within model to account for the association of the random effects with the exposure and the cluster population sizes. Our approach represents a new way of thinking about population-average effects with mixed effects models. We illustrate with two examples concerning, respectively,alcohol consumption and antibiotic use.

✉ brumback@ufl.edu

ASSESSING INDIVIDUAL AND DISSEMINATED CAUSAL PACKAGE EFFECTS IN NETWORK HIV TREATMENT AND PREVENTION TRIALS

Ashley Buchanan*, *University of Rhode Island*

Donna Spiegelman, *Harvard School of Public Health*

Sten Vermund, *Yale School of Public Health*

Samuel Friedman, *National Development and Research Institutes, Inc.*

Judith Lok, *Harvard School of Public Health*

Evaluation of packages of prevention interventions for HIV and other infectious diseases are needed because many interventions offer partial protection. We propose an approach to evaluate the causal effects of package com-

ponents in a single study with a multifaceted intervention. Some participants randomized to the intervention are exposed directly but others only indirectly. The individual effect is that among directly exposed participants beyond being in an intervention network; the disseminated effect is that among participants not directly exposed. We estimated individual and disseminated package component effects in HIV Prevention Trials Network 037, a Phase III network-randomized HIV prevention trial among persons who inject drugs and their risk networks. The index participant in an intervention network received an initial and booster peer education intervention and all participants were followed to ascertain risk behaviors. We used marginal structural models to adjust for time-varying confounding. These methods will be useful for evaluation of the causal effects of package interventions in a single study with network features.

✉ buchanan@uri.edu

A "POTENTIAL OUTCOMES" APPROACH TO ACCOUNT FOR MEASUREMENT ERROR IN MARGINAL STRUCTURAL MODELS

Jessie K. Edwards*, *University of North Carolina, Chapel Hill*

Marginal structural models are important tools for observational studies, but these models typically assume that variables are measured without error. This work demonstrates how bias due to measurement error can be described in terms of potential outcomes and describes a method to account for nondifferential or differential measurement error in a marginal structural model. We illustrate the proposed method estimating the joint effects of antiretroviral therapy initiation and smoking on all-cause mortality in a cohort of 12,290 patients with HIV followed for 5 years. Smoking status in the total population was likely measured with error, but a subset of 3686 patients who reported smoking status on separate questionnaires composed an internal validation subgroup. We compared a standard joint marginal structural model fit using inverse probability weights to a model that accounted for misclassification of smoking status using an imputation approach.

✉ jessedwards@unc.edu

73. STATISTICAL METHODS IN SINGLE-CELL GENOMICS

REMOVING UNWANTED VARIATION USING BOTH CONTROL AND TARGET GENES IN SINGLE CELL RNA SEQUENCING STUDIES

Mengjie Chen*, *University of Chicago*

Xiang Zhou, *University of Michigan*

Single cell RNA sequencing (scRNAseq) technique is becoming increasingly popular for unbiased and high-resolutional transcriptome analysis of heterogeneous cell populations. Despite its many advantages, scRNAseq, like any other genomic sequencing technique, is susceptible to the influence of confounding effects. Controlling for confounding effects in scRNAseq data is thus a crucial step for proper data normalization and accurate downstream analysis. Several recent methodological studies have demonstrated the use of control genes (including spike-ins) for controlling for confounding effects in scRNAseq studies. However, these methods can be suboptimal as they ignore the rich information contained in the target genes. Here, we develop an alternative statistical method, which we refer to as scPLS, for more accurate inference of confounding effects. Our method models control and target genes jointly to better infer and control for confounding effects. With simulations and studies, we show the effectiveness of scPLS in removing technical confounding effects as well as for removing cell cycle effects.

✉ mengjiechen@uchicago.edu

CELL SIMILARITY MEASURES FOR IDENTIFYING CELL SUBPOPULATIONS FROM SINGLE-CELL RNA-Seq DATA

Haiyan Huang*, *University of California, Berkeley*

One goal of single cell RNA-sequencing (scRNA-seq) is to expose possible heterogeneity within cell populations due to meaningful, biological variation. Examining cell-to-cell heterogeneity, and further, identifying subpopulations of cells based on scRNA-seq data has been of common interest

in life science research. A key component to successfully identifying cell subpopulations (or clustering cells) is the (dis)similarity measure used to group the cells. In this talk, I introduce a novel measure to assess cell-to-cell similarity using scRNA-seq data. This new measure incorporates information from all cells when evaluating the similarity between any two cells, a characteristic not commonly found in existing (dis)similarity measures. This property is advantageous for two reasons: (a) borrowing information from cells of different subpopulations allows for the investigation of pair-wise cell relationships from a global perspective, and (b) information from other cells of the same subpopulation could help to ensure a robust relationship assessment.

✉ hhuang@stat.berkeley.edu

SINGLE-CELL ATAC-Seq SIGNAL EXTRACTION AND ENHANCEMENT

Hongkai Ji*, *Johns Hopkins Bloomberg School of Public Health*

Zhicheng Ji, *Johns Hopkins Bloomberg School of Public Health*

Weiqiang Zhou, *Johns Hopkins Bloomberg School of Public Health*

Single-cell assay of transposase-accessible chromatin followed by sequencing (scATAC-seq) is an emerging new technology for studying gene regulation. Unlike the conventional ChIP-seq, DNase-seq and ATAC-seq technologies which measure average behavior of a cell population, scATAC-seq measures regulatory element activities within each individual cell, thereby allowing one to examine the heterogeneity of a cell population. Analyzing scATAC-seq data is challenging because the data are highly sparse and discrete. We present a statistical model to effectively extract signals from the noisy scATAC-seq data. Our method leverages information in massive amounts of publicly available DNase-seq data to enhance the scATAC-seq signal. We demonstrate through real data analyses that this approach substantially improves the accuracy for reconstructing genome-wide regulatory element activities.

✉ hji@jhu.edu

NORMALIZATION AND REPRODUCIBILITY IN SINGLE CELL RNA-Seq

Zhijin Wu*, *Brown University*

Single cell RNA-seq (scRNA-seq) enables the transcriptomic profiling at individual cell level. This new level of resolution reveals inter-cellular transcriptomic heterogeneity and brings new promises to the understanding of transcriptional regulation mechanism. Similar to data from other high-throughput technologies, scRNA-seq data are affected by substantial technical and biological artifacts, maybe more so due to the low amount of starting materials and more complex sample preparation. With high heterogeneity expected between cells, normalization faces new challenge because typical assumptions made for bulk RNA samples no longer hold. Yet it is still a necessary step for proper comparison and to ensure reproducibility between studies. We discuss the unique challenges in normalization of scRNA-seq data and the impact of different strategies in normalization on the analysis of scRNA-seq data. We present a probabilistic model of sequencing counts that well explains the characteristics of single cell RNA-seq data and an adaptive normalization procedure that is robust to the bursting nature of expression in many genes.

✉ zhijin_wu@brown.edu

74. FUNCTIONAL DATA ANALYSIS

STATISTICAL MODELS IN SENSORY QUALITY CONTROL: THE CASE OF BOAR TAINT

Jan Gertheiss*, *Clausthal University of Technology*

Johanna Mörlein, *Georg August University Göttingen*

Lisa Meier-Dinkel, *Georg August University Göttingen*

Daniel Mörlein, *Georg August University Göttingen*

The rearing of entire male pigs is avoided in most countries because of its association with so-called boar taint. An estimated number of more than 100 million male piglets

is therefore surgically castrated per year in the EU. This practice, however, has been proven painful for the animals, and increasing public demand for improved animal welfare made the European pork production chain stakeholders declare the ban of surgical castration by 2018. Despite some advantages of fattening boars, the ban of castration is linked to the risk of impaired consumer acceptance of pork as a result of boar taint, which is mainly caused by two malodorous volatile substances: androstenone and skatole. In the talk, we will consider data from an experimental study where fat samples of more than a thousand pig carcasses were collected and subjected to a thorough sensory evaluation and quantification using a panel of 10 trained assessors. We will discuss various statistical models for analyzing and quantifying the influence of androstenone and skatole on the olfactory perception of boar taint, including parametric, nonparametric and ordinal regression models.

✉ jan.gertheiss@tu-clausthal.de

PRINCIPAL COMPONENT ANALYSIS FOR SPATIALLY DEPENDENT FUNCTIONAL DATA

Haozhe Zhang*, *Iowa State University*

Yehua Li, *Iowa State University*

We consider spatially dependent functional data collected under a goestatistics setting, where locations are sampled from a spatial point process and a random function is observed at each location. Observations on each function are made on discrete time points and contaminated with nugget effects and measurement errors. The error process at each location is modeled as a non-stationary temporal process rather than white noise. Under the assumption of spatial isotropy, we propose a tensor product spline estimator for the spatio-temporal covariance function. If a coregionalization covariance structure is further assumed, we propose a new functional principal component analysis method that borrow information from neighboring functions. Under a unified framework for both sparse and dense functional data, where the number of observations per curve is allowed to be of any rate relative to the number of functions, we develop the asymptotic convergence rates for the proposed estimators. The proposed methods are illustrated by simulation studies and a real data application.

✉ haozhe@iastate.edu

NON-PARAMETRIC FUNCTIONAL ASSOCIATION TEST

Sneha Jadhav*, *Yale University*

Shuangge Ma, *Yale University*

In this paper, we develop a non-parametric method to test for association between a scalar variable and a functional variable. We propose a functional U-statistic and establish asymptotic distribution of this statistic under the null hypothesis of no association. This result is used to construct the association test. In the simulation section we first demonstrate the need for a non-parametric functional test. We use simulations to study some properties of this test, explore it's applicability to association studies in sequencing data and compare it's performance with that of sequence kernel association test. We also present a modification of this test to accommodate covariates and study it's performance in the simulations. Finally, we present a real data application.

✉ snehaj19@gmail.com

NON-PARAMETRIC FUNCTIONAL ASSOCIATION TEST

Sneha Jadhav*, *Yale University*

Shuangge Ma, *Yale University*

In this paper, we develop a non-parametric method to test for association between a scalar variable and a functional variable. We propose a functional U-statistic and establish asymptotic distribution of this statistic under the null hypothesis of no association. This result is used to construct the association test. In the simulation section we first demonstrate the need for a non-parametric functional test.

We use simulations to study some properties of this test, explore it's applicability to association studies in sequencing data and compare it's performance with that of sequence kernel association test. We also present a modification of this test to accommodate covariates and study it's performance in the simulations. Finally, we present a real data application.

✉ snehaj19@gmail.com

REGISTRATION FOR EXPONENTIAL FAMILY FUNCTIONAL DATA

Julia Wrobel* •, *Columbia University*

Jeff Goldsmith, *Columbia University*

We introduce a novel method for separating amplitude and phase variability in exponential family functional data. Our method alternates between two steps: the first uses generalized functional principal components analysis to calculate template functions, and the second estimates warping functions that map observed curves to templates. Existing approaches to registration have focused on continuous functional observations, and the few approaches for discrete functional data require pre-smoothing. In contrast, we focus on the likelihood of the observed data and avoid the need for preprocessing, and we implement both steps of our algorithm in a computationally efficient way. Our motivation comes from the Baltimore Longitudinal Study on Aging, in which accelerometer data provides insights into the timing of sedentary behavior. We analyze binary functional data with observations each minute over 24 hours for 579 participants, where values represent activity and inactivity. Diurnal patterns of activity are obscured due to misalignment in the original data but are clear after curves are aligned. Simulations designed to mimic our application outperform competing approaches.

✉ jw3134@cumc.columbia.edu

OPTIMAL DESIGN FOR CLASSIFICATION OF FUNCTIONAL DATA

Cai Li*, *North Carolina State University*

Luo Xiao, *North Carolina State University*

We study the design problem for optimal classification of functional data. The goal is to select sampling time points so that functional data observed at these time points can be classified as accurately as possible. We propose optimal designs that are applicable for a pilot study with either dense or sparse functional data. Using linear discriminant analysis, we formulate our design objectives as explicit functions of the sampling points. We study the theoretical properties of the proposed design objectives and provide a practical implementation. The performance of the proposed design is assessed through simulations and real data applications.

✉ cli9@ncsu.edu

75. HIGH DIMENSIONAL DATA ANALYSIS

ROBUST ANALYSIS OF HIGH DIMENSIONAL DATA

Quefeng Li*, *University of North Carolina, Chapel Hill*

Marco Avella-Medina, *Massachusetts Institute of Technology*

Jianqing Fan, *Princeton University*

Heather Batty, *Imperial College London*

In the last decade, many new statisical tools have been developed to handle the large-p-small-n problem. However, most of these tools rely on the assumption that the underlying distribution is light-tailed (i.e. close to the Gaussian distribution). In the high dimensional setting, when many variables are involved, such an assumption is often too strong. In data collected from the real world, such as genomic data and neuroimaging data, we often observe outliers, skewness, and other aspects that clearly indicate that the underlying distribution is very different from Gaussian. Therefore, it is important to develop robust methods with guaranteed statistical

properties for analyzing data that are collected from heavy-tailed distributions. In this talk, we will discuss the robust estimation of covariance/precision matrix and the robust linear regression under the high dimensional setting.

✉ quefeng@email.unc.edu

A DISTRIBUTED AND INTEGRATED METHOD OF MOMENTS FOR HIGH-DIMENSIONAL CORRELATED DATA ANALYSIS

Emily C. Hector* •, *University of Michigan*

Peter X. K. Song, *University of Michigan*

This paper is motivated by a regression analysis of electroencephalography (EEG) data with high-dimensional correlated responses with multi-level nested correlations. We develop a divide-and-conquer procedure implemented in a distributed and parallelized computational scheme for statistical estimation and inference of regression parameters. The computational bottleneck associated with high-dimensional likelihoods prevents the scalability of existing methods. The proposed method addresses this by dividing responses into subvectors to be analyzed in parallel using composite likelihood. Theoretical challenges related to combining results from dependent data are overcome in a statistically efficient way with a meta-estimator derived from Hansen's generalized method of moments. We provide a theoretical framework for efficient estimation, inference, and goodness-of-fit tests, and develop an R package. We illustrate our method's performance with simulations and the analysis of the EEG data, and find that iron deficiency is significantly associated with two electrical potentials related to auditory recognition memory in the left parietal-occipital region of the brain.

✉ ehector@umich.edu

TROPICAL PRINCIPAL COMPONENT ANALYSIS AND ITS APPLICATION TO PHYLOGENETICS

Xu Zhang*, *University of Kentucky*

Ruriko Yoshida, *Naval Postgraduate School*

Leon Zhang, *University of California, Berkeley*

Principal component analysis is a widely-used method for the dimensionality reduction of a given data set in a high-dimensional Euclidean space. Here we define and analyze two analogues of principal component analysis in the setting of tropical geometry. In one approach, we study the Stiefel tropical linear space of fixed dimension closest to the data points in the tropical projective torus; in the other approach, we consider the tropical polytope with a fixed number of vertices closest to the data points. We then give approximative algorithms for both approaches and apply them to phylogenetics, testing the methods on simulated phylogenetic data and on an empirical dataset of Apicomplexa genomes.

✉ xzh323@g.uky.edu

USING SUFFICIENT DIRECTION FACTOR MODEL TO ANALYZE BREAST CANCER PATHWAY EFFECTS

Seungchul Baek*, *University of South Carolina*

Yen-Yi Ho, *University of South Carolina*

Yanyuan Ma, *The Pennsylvania State University*

We propose a new analysis paradigm for breast cancer survival data with gene expression. In the first stage, under ultra high dimensional covariates, we estimate factor and loading matrices based on several cancer types. At the same time, an additional sparse condition on loading matrix can be imposed according to prior knowledge. In the second stage, we then employ the general index model for survival data, which is developed by a semiparametric regime. We show the performance of finite samples by conducting simulations and find that the modeling we propose works well in this complicated data structure. In the data analysis, we provide some interpretations about pathways effects and genes for breast cancer.

✉ sbaek@email.sc.edu

INFERENCE FOR HIGH-DIMENSIONAL LINEAR MEDIATION ANALYSIS MODELS IN GENOMICS

Ruixuan Zhou*, *University of Illinois at Urbana-Champaign*

Liewei Wang, *Mayo Clinic*

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

We propose two new inference procedures for high-dimensional linear mediation analysis models, where the number of potential mediators can be much larger than the sample size. We first propose estimators for direct and indirect effects under incomplete mediation and prove their consistency and asymptotic normality. We next consider the complete mediation setting where the direct effect is known to be absent. We propose an estimator for the indirect effect and establish its consistency and asymptotic normality. Furthermore, we prove that our approach gives a more powerful test compared to directly testing for the total effect, which equals the indirect effect under complete mediation. We confirm our theoretical results in simulations. In an integrative analysis of gene expression and genotype data from a pharmacogenomic study of drug response in human lymphoblastoid cell lines, we use our first method to study the direct and indirect effects of coding variants on drug responses. We use our second method to identify a genome-wide significant noncoding variant that was not detected using standard genome-wide association study analysis methods.

✉ rzhou14@illinois.edu

MULTIVARIATE DENSITY ESTIMATION VIA MINIMAL SPANNING TREE AND DISCRETE CONVOLUTION

Zhipeng Wang*, *Rice University*

David Scott, *Rice University*

Density estimation is the building block for a variety of tasks in statistical inference and machine learning, including anomaly detections, classifications, clustering and image analysis etc. Conventional nonparametric density estimators such as the Kernel Density Estimator and Histograms etc. cannot provide reliable density estimates for high-dimensional data due to the fact that the number of data points needed grows exponentially with the number of dimensions. In this work, we proposed a novel method using Minimal Spanning Tree (MST), a widely-adopted algorithm in computer science, to form a parsimonious representation of the high-dimensional data. Based on the MST we developed a greedy algorithm to partition the tree to come up with clusters. We then further utilize the centroids and the sizes of clusters to perform discrete convolution over the entire data domain via kernel smoothing. The nonparametric density estimator developed by this work provides an efficient, adaptive and robust density estimate for high-dimensional data and it is relatively insensitive to noise.

✉ Zhipeng.Wang@alumni.rice.edu

IMPUTATION USING LINKED MATRIX FACTORIZATION

Michael J. O'Connell*, *University of Minnesota*

Eric F. Lock, *University of Minnesota*

Several recent methods address the dimension reduction and decomposition of linked high-content data matrices. Typically, these methods consider one dimension, rows or columns, that is shared among the matrices. This shared dimension may represent common features measured for different sample sets or a common sample set with features from different platforms. We discuss an approach for simultaneous horizontal and vertical integration, Linked Matrix Factorization (LMF), for the case where some matrices share rows and some share columns. Our motivating application is a cytotoxicity study with genomic and molecular chemical attribute data. The toxicity matrix (cell lines x chemicals) shares samples with a genotype matrix (cell lines x SNPs) and features with a molecular attribute matrix (chemicals

x attributes). LMF gives a unified low-rank factorization of these three matrices, which allows for the decomposition of systematic variation that is shared and that is specific to each matrix. We use this for the imputation of missing data even when entire rows or columns are missing. We also introduce two methods for estimating the ranks of the joint and individual structure.

✉ oconn725@umn.edu

76. METHODS FOR CATEGORICAL AND ORDINAL DATA

BAYESIAN TESTING FOR INDEPENDENCE OF TWO CATEGORICAL VARIABLES WITH COVARIATES UNDER CLUSTER SAMPLING

Dilli Bhatta*, *University of South Carolina Upstate*

We consider Bayesian testing for independence of two categorical variables with covariates for a two-stage cluster sample. This is a difficult problem because we have a complex sample, not a simple random sample. Our approach 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. We show that if there is strong association between two categorical variables, there is no significant difference between the tests with and without the covariates. We also performed a simulation study to further understand the effect of covariates in various situations. We found that for borderline cases (moderate association between the two categorical variables), there are noticeable differences in the test with and without covariates.

✉ dbhatta@uscupstate.edu

A REVIEW AND CRITIQUE OF STATISTICAL METHODS FOR THE ANALYSIS OF VENTILATOR-FREE DAYS

Charity J. Morgan*, *University of Alabama at Birmingham*

Yuliang Liu, *University of Alabama at Birmingham*

The number of ventilator-free days (VFDs) is a common endpoint for clinical trials assessing lung injury or acute respiratory distress syndrome. A patient's VFD is defined as the total number of days the patient is both alive and free of mechanical ventilation; patients who die during observation are assigned a VFD of zero. Despite usually being both truncated and zero-inflated, VFDs are often analyzed using statistical methods that assume normally distributed data. While more sophisticated data analytic approaches, such as nonparametric and competing risk analyses, have been proposed, their use is not yet widespread in the critical care literature and their applicability to this endpoint remains the source of debate. We review the existing critical care literature and compare these methods via simulations and real data examples.

✉ cjmorgan@uab.edu

ONLINE ROBUST FISHER DISCRIMINANT ANALYSIS

Hsin-Hsiung Huang*, *University of Central Florida*

Teng Zhang, *University of Central Florida*

We introduce an algorithm which solve Fisher linear discriminant analysis (LDA) by iterations for streamline data and the results are robust to outliers. The proposed iterative robust LDA combines the merits of iterative updating and robust LDA. It inherits good properties from these two ideas for reducing the time complexity, space complexity, and the influence of these outliers on estimating the principal directions. In the asymptotic stability analysis, we also show that our online robust LDA converges to the weighted kernel principal kernel components from the batch robust LDA given

good initial values. Experimental results are presented to confirm that our online robust LDA is effective and efficient. The proposed method is able to deal with high dimensional data ($p>n$) and is robust against outliers, so that it can be applied to classify microarray gene expression datasets.

✉ hsin.huang@ucf.edu

BAYESIAN ORDINAL RESPONSE MODELS FOR IDENTIFYING MOLECULAR MECHANISMS IN THE PROGRESSION TO CERVICAL CANCER

Kellie J. Archer*, *The Ohio State University*

Yiran Zhang, *The Ohio State University*

Qing Zhou, *U.S. Food and Drug Administration*

Pathological evaluations are frequently reported on an ordinal scale. Moreover, diseases may progress from less to more advanced stages. For example, cervical cancer due to HPV infection progresses from normal epithelium, to low-grade squamous intraepithelial lesions, to high-grade squamous intraepithelial lesions (HSIL), and then to invasive carcinoma. To elucidate molecular mechanisms associated with disease progression, genomic characteristics from samples procured from these different tissue types were assayed using a high-throughput platform. Motivated by Park and Casella's (2008) Bayesian LASSO, we developed a penalized ordinal Bayesian model that incorporates a penalty term so that a parsimonious model can be obtained. Through simulation studies, we investigated different formulations of threshold parameters and their priors and compared our penalized ordinal Bayesian model to penalized ordinal response models fit using frequentist-based approaches. We applied our penalized ordinal Bayesian methodology to identify molecular features associated with normal squamous cervical epithelial samples, HSIL, and invasive squamous cell carcinomas of the cervix.

✉ archer.43@osu.edu

SAMPLE SIZE ESTIMATION FOR MARGINALIZED ZERO-INFLATED COUNT REGRESSION MODELS

Leann Long*, *University of Alabama at Birmingham*

Dustin Long, *University of Alabama at Birmingham*

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

Recently, the marginalized zero-inflated (MZI) Poisson and MZI negative binomial models have been proposed to provide direct overall exposure effects estimation rather than the latent class interpretations provided by the traditional zero-inflated framework. We briefly discuss the motivation and potential advantages of this MZI methodology for count data with excess zeroes in health research. Also, we examine sample size calculations for the MZI methods for testing marginal means of two groups of independent observations where zero-inflated counts are expected. Currently available methods focus on Poisson or traditional zero-inflated Poisson regression model for sample size calculation. We compare sample size calculations from the Poisson and MZI regression models for efficiency considerations. Through the derivation and assessment of these sample size calculations, the convenient marginal mean interpretations of MZI methods can be utilized in the planning of future scientific work.

✉ leannl@uab.edu

77. MULTIVARIATE METHODS

REGRESSION TREES AND ENSEMBLE METHODS FOR MULTIVARIATE OUTCOMES

Evan L. Reynolds*, *University of Michigan*

Mousumi Banerjee, *University of Michigan*

Tree-based methods have become one of the most flexible, intuitive, and powerful data analytic tools for exploring complex data structures. The best documented, and arguably most popular uses of tree-based methods are in biomedical research, where multivariate outcomes occur commonly (e.g. diastolic and systolic blood pressure, periodon-

tal measures in dental health studies, and nerve density measures in studies of neuropathy). Existing tree-based methods for multivariate outcomes do not appropriately take into account the correlation that exists in the outcomes. In this paper, we develop two goodness of split measures for multivariate tree building for continuous outcomes. The proposed measures summarize specific aspects of the variance-covariance matrix at each potential split. Additionally, to enhance prediction accuracy we extend the single multivariate regression tree to an ensemble of trees. Extensive simulations are presented to examine the properties of our goodness of fit measures. Finally, the proposed methods are illustrated using two clinical datasets of neuropathy and pediatric cardiac surgery.

✉ evanlr@umich.edu

SimMultiCorrData: AN R PACKAGE FOR SIMULATION OF CORRELATED VARIABLES OF MULTIPLE DATA TYPES

Allison C. Fialkowski*, *University of Alabama at Birmingham*

Hemant K. Tiwari, University of Alabama at Birmingham

There is a dearth of R packages capable of simulating correlated data sets of multiple variable types with high precision and valid pdfs. Both are required for statistical model comparisons or power analyses. The package SimMultiCorrData generates correlated continuous, ordinal, Poisson, and Negative Binomial variables that mimic real-world data sets. Continuous variables are simulated using either Fleishmans 3rd-order or Headricks 5th-order power method transformation. The fifth-order PMT permits control over higher moments and generation of more kurtoses and valid pdfs. Two simulation pathways provide distinct methods for calculating the correlations. Additional functions calculate cumulants, determine correlation and lower kurtosis boundaries, check for a valid pdf, and summarize or graph the simulated variables. Examples contrast the two simulation functions, demonstrate the optional error loop to minimize correlation error, and compare this package to Demirtas et al.s (2017) PoisBinOrdNonNor. SimMultiCorrData provides the 1st R implementation of the 5th-order PMT and enhances existing correlated data simulation packages by allowing Negative Binomial variables.

✉ allijazz@uab.edu

SPARSE MULTIPLE CO-INERTIA ANALYSIS WITH APPLICATIONS TO 'OMICS DATA

Eun Jeong Min*, *University of Pennsylvania*

Qi Long, *University of Pennsylvania*

Multiple co-inertia analysis (mCIA) is a multivariate statistical analysis method that can access relationships and trends between multiple sets of data. While originally mCIA has been widely used in ecology, more recently it has been used for integrative analysis of multiple high-dimensional omics datasets. The estimated loading vectors of classical mCIA are not sparse, which may present challenges in interpreting analysis results particularly when analyzing omics data. We propose a sparse mCIA (smCIA) method that imposes sparsity in estimated loading vectors via regularization. The resulting sparse loading vectors can provide insights on important omics features that contribute to and account for most of the correlations between multiple -omics datasets. Synthetic data and real omics data are used to demonstrate the superior performance of our approach to existing mCIA methods.

✉ mineunj@pennmedicine.upenn.edu

SMALL SPHERE DISTRIBUTIONS FOR DIRECTIONAL DATA WITH APPLICATION TO MEDICAL IMAGING

Byungwon Kim*, *University of Pittsburgh*

Stephan Huckemann, *University of Göttingen*

Jorn Schulz, *University of Stavanger*

Sungkyu Jung, *University of Pittsburgh*

We propose new small-sphere distributional families for modeling multivariate directional data. In a special case of univariate directions, the new densities model random

directions on the unit sphere with a tendency to vary along a small circle on the sphere, and with a unique mode on the small circle. The proposed multivariate densities enable us to model association among multivariate directions, and are useful in medical imaging, where multivariate directions are used to represent shape and shape changes of 3-dimensional objects. When the underlying objects are rotationally deformed under noise, for instance, twisted and/or bend, corresponding directions tend to follow the proposed small-sphere distributions. The proposed models have several advantages over other methods analyzing small-circle-concentrated data, including inference procedures on the association and small-circle fitting. We demonstrate the use of the proposed multivariate small-sphere distributions in analysis of skeletally-represented object shapes.

✉ byk4@pitt.edu

SUPER-DELTA: A NEW APPROACH THAT COMBINES GENE EXPRESSION DATA NORMALIZATION AND DIFFERENTIAL EXPRESSION ANALYSIS

Yuhang Liu*, *Florida State University*

Jinfeng Zhang, *Florida State University*

Xing Qiu, *University of Rochester*

In this study we propose a new differential expression analysis pipeline, dubbed as super-delta. It consists of a robust multivariate extension of global normalization designed to minimize the bias introduced by DEGs, suitably paired by a modified t-test based on asymptotic theory for hypothesis testing. We first compared super-delta with commonly used normalization methods: global, median-IQR, quantile, and cyclic-loess normalization in simulation studies. Super-delta was shown to have better statistical power with tighter type I error control than its competitors. In many cases, its performance is close to that of using oracle datasets without technical noise. We then applied all methods to a dataset of breast cancer patients receiving neoadjuvant chemotherapy. While there is a substantial overlap of DEGs identified by all methods, super-delta was able to identify comparatively more DEGs. As a new pipeline, super-delta provides new insights to the area of differential expression analysis. Its decent performance is supported by solid theoretical foundation and demonstrated by both real data and simulation analysis. Its multi-group extension is under active development.

✉ fhlsjs@gmail.com

EXCEEDANCE PROBABILITIES FOR EXCHANGEABLE RANDOM VARIABLES

Satish Iyengar*, *University of Pittsburgh*

Burcin Simsek, *Bristol-Myers Squibb*

The impact of correlated inputs to neurons is of interest because in vivo recordings in the rat somatosensory cortex indicate that such correlation is present in both spontaneous and stimulated neural activity. A special case used in computational models of neural activity assumes that the inputs are exchangeable, and that a neuron spikes when a certain number of the inputs exceed a certain threshold. In this paper, we study exceedance probability distributions: in particular, we give conditions under which they are unimodal and prove certain majorization results as the correlation coefficient varies. We also give asymptotic approximations that are useful for studying large numbers of inputs.

✉ ssi@pitt.edu

78. SMART DESIGNS AND DYNAMIC TREATMENT REGIMENS

A BAYESIAN ANALYSIS OF SMALL N SEQUENTIAL MULTIPLE ASSIGNMENT RANDOMIZED TRIALS (snSMARTs)

Boxian Wei*, *University of Michigan*

Thomas M. Braun, *University of Michigan*

Roy N. Tamura, *University of South Florida*

Kelley M. Kidwell, *University of Michigan*

Designing clinical trials for rare diseases is challenging because of the limited number of patients. A suggested design is the small-n Sequential Multiple Assignment Randomized Trial (snSMART), in which patients are first randomized to one of multiple treatments (stage 1). Patients who respond continue the same treatment for another stage, while those who fail to respond are re-randomized to one of the remaining treatments (stage 2). We propose a Bayesian approach to compare the efficacy between treatments allowing for borrowing of information across both stages. Via simulation, we compare the bias, root mean-square error (rMSE), width and coverage rate of 95% confidence/credible interval (CI) of estimators from our approach to estimators produced from (a) standard approaches that only use the data from stage 1, and (b) a log-Poisson model using data from both stages whose parameters are estimated via generalized estimating equations. The rMSE and width of 95% CIs of our estimators are smaller than the other approaches in realistic settings, so that the collection and use of stage 2 data in snSMARTs provide improved inference for treatments of rare diseases.

✉ boxian@umich.edu

EVALUATING THE EFFECTS OF MISCLASSIFICATION IN SEQUENTIAL MULTIPLE ASSIGNMENT RANDOMIZED TRIALS (SMART)

Jun He*, *Virginia Commonwealth University*

Donna McClish, *Virginia Commonwealth University*

Roy Sabo, *Virginia Commonwealth University*

SMART designs tailor individual treatment by re-randomizing patients to subsequent therapies based on their response to initial treatment. However, patients' response could be misclassified. They could be allocated to inappropriate second-stage treatments; statistical analysis could be affected. Thus, we aim to evaluate the effect of misclassification on SMART designs with respect to bias of means and variances, and the power of outcome comparisons. We focus on comparing dynamic treatment regimens, a set of decision rules used to choose between treatments for individual patients based on respond to initial treatment. Assuming continuous responses, equal randomization, and equal variances, we derived formulas to analytically investigate bias and power as functions of sensitivity, specificity, true response rate, and effect size. The results show that misclassification produces biased estimates of mean and variance. Power is usually reduced; the relationship between power and sensitivity/specificity can be non-monotonic. These findings show that misclassification can adversely affect SMART designs, and suggest the development of methods to minimize these effects.

✉ jhe3@vcu.edu

POWER ANALYSIS IN A SMART DESIGN: SAMPLE SIZE ESTIMATION FOR DETERMINING THE BEST DYNAMIC TREATMENT REGIME

William J. Artman*, *University of Rochester*

Tianshuang Wu, *AbbVie*

Ashkan Ertefaie, *University of Rochester*

Sequential, multiple assignment, randomized trial (SMART) designs have gained considerable attention in the field of precision medicine by providing a cost effective and empirically rigorous platform for comparing sequences of treatments tailored to the individual patient, i.e., dynamic treatment regime (DTR). The construction of evidence-based DTRs promises an alternative to adhoc one-size-fits-all decisions pervasive in patient care. However, the advent of SMART designs poses substantial statistical challenges in performing power analyses due to the complex correlation structure between the DTRs embedded in the design. Since the main goal of SMARTs is to construct an optimal DTR, investigators are interested in sizing such trials based on the ability to screen out DTRs inferior to the optimal DTR by a given amount which cannot be done using existing methods. We fill this gap by developing a rigorous power analysis framework that leverages multiple comparisons with the best methodology. We demonstrate the

validity of our method through extensive simulation studies and illustrate its application using the Extending Treatment Effectiveness of Naltrexone SMART study data.

✉ William_Artman@URMC.Rochester.edu

DYNAMIC TREATMENT REGIMES WITH SURVIVAL OUTCOMES

Gabrielle Simoneau*, *McGill University*

Robert W. Platt, *McGill University*

Erica E.M. Moodie, *McGill University*

A dynamic treatment regime (DTR) is a set of decision rules to be applied across multiple stages of treatments. The decisions are tailored to individuals, by inputting an individual's observed characteristics and outputting a treatment decision at each stage for that individual. Of interest is the identification of an optimal DTR, that is, the sequence of treatment decisions that yields the best expected outcome for a population of 'similar' individuals. Unlike uncensored continuous or dichotomous outcomes, there exist only a few statistical methods that consider the problem of identifying an optimal DTR with time-to-event data subject to right censoring. I propose to extend a theoretically robust and easily implementable method for estimating an optimal DTR, dynamic weighted ordinary least squares (dWOLS), to accommodate time-to-event data. I will explain the statistical methodology behind dWOLS for continuous outcomes, and provide conceptual and theoretical details on the proposed modifications to extend the method to time-to-event data. I will show that, as for dWOLS, the proposed extension is doubly-robust, easy to understand and easily applicable.

✉ gabrielle.simoneau@mail.mcgill.ca

79. SURVIVAL ANALYSIS AND SEMI- AND NON-PARAMETRIC MODELS

MARTINGALE-BASED OMNIBUS TESTS FOR SEMIPARAMETRIC TRANSFORMATION MODEL WITH CENSORED DATA

Soutrik Mandal*, *Texas A&M University*

Suojin Wang, *Texas A&M University*

Samiran Sinha, *Texas A&M University*

Censored time-to-event data are often analyzed using semiparametric linear transformation models which contain popular models like the Cox proportional hazards model and the proportional odds model as special cases. A misspecified model leads to invalid inference. We propose a new class of omnibus supremum tests derived from martingale-based residuals to test goodness-of-fit of the assumed model. We derive the analytical expression of the test statistics under the null hypothesis and assess it through a Monte Carlo method. The superiority of our tests over existing methods is demonstrated through simulation studies and real data examples.

✉ smandal@stat.tamu.edu

PENALIZED ESTIMATION OF GENERALIZED ADDITIVE COX MODEL FOR INTERVAL-CENSORED DATA

Yan Liu*, *University of Nevada, Reno*

Minggen Lu, *University of Nevada, Reno*

Christopher McMahan, *Clemson University*

In this work, we propose a generalized additive Cox proportional hazard model for interval-censored data. In the proposed model, unknown functions are approximated through the use of smoothing splines. To obtain the maximum likelihood estimates (MLE) of regression parameters and spline coefficients, an accelerated expectation-maximization algorithm is used. Under standard regularity conditions, the asymptotic normality and efficiency of the

MLE of regression parameters is established, and the nonparametric estimator of the unknown functions is shown to achieve the optimal rate of convergence. Through extensive Monte Carlo simulation studies, it is shown that the proposed approach can accurately and efficiently estimate all unknown model parameters. The proposed approach is further illustrated using data from a large population-based randomized trial designed and sponsored by the United States National Cancer Institute.

✉ yliu23@unr.edu

RESTRICTED MEAN SURVIVAL TIME FOR RIGHT-CENSORED DATA WITH BIASED SAMPLING

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

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

Yu Shen, University of Texas MD Anderson Cancer Center

In clinical studies with time-to-event outcomes, the restricted mean survival time (RMST) has attracted substantial attention as a summary measurement for its straightforward clinical interpretation. When the data are subject to biased sampling, which is frequently encountered in observational cohort studies, existing methods to estimate the RMST are not applicable. In this paper, we consider nonparametric and semiparametric regression methods to estimate the RMST under the setting of length-biased sampling. To assess the covariate effects on the RMST, a semiparametric regression model that directly relates the covariates and the RMST is assumed. Based on the model, we develop unbiased estimating equations to obtain consistent estimators of covariate effects by properly adjusting for informative censoring and length bias. In addition, we further extend the methods to account for general left-truncation. We investigate the finite sample performance through simulations and illustrate the methods by analyzing a prevalent cohort study of dementia in Canada.

✉ clee9@mdanderson.org

ON THE SURVIVOR CUMULATIVE INCIDENCE FUNCTION OF RECURRENT EVENTS

Lu Mao*, *University of Wisconsin, Madison*

Assessment of recurrent endpoints in the presence of a terminal event, such as patient death, has always been a challenging, and sometimes controversial, issue in biomedical studies. Statistical analysis based on the cumulative incidence of recurrent events may lead to dubious conclusions because the incidence is highly susceptible to changes in the death rate. We propose to analyze death-terminated recurrent event processes by the survivor cumulative incidence function, which is the average cumulative number of recurrent events up to certain time point among those who have survived to that point. We construct a naive nonparametric estimator for the survivor cumulative incidence function and use semiparametric theory to improve its statistical efficiency without compromising robustness. A class of hypothesis tests is developed to compare the function between groups. Extensive simulations studies demonstrate that the proposed procedures perform satisfactorily in realistic sample sizes. A dataset from a major cardiovascular study is analyzed to illustrate our methods.

✉ lmao@biostat.wisc.edu

SOME ASYMPTOTIC RESULTS FOR SURVIVAL TREES AND FORESTS

Yifan Cui*, *University of North Carolina, Chapel Hill*

Ruoqing Zhu, *University of Illinois at Urbana-Champaign*

Mai Zhou, *University of Kentucky*

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

This paper develops a theoretical framework and asymptotic results for survival tree and forest models under right censoring. We first investigate the method from the aspect of splitting rules, where the survival curves of the two potential child nodes are calculated and compared. We show that existing approaches lead to a potentially biased estimation

of the within-node survival and cause non-optimal selection of the splitting rules. This bias is due to the censoring distribution and the non i.i.d. sample structure within each node. Based on this observation, we develop an adaptive concentration bound result for both tree and forest versions of the survival tree models. The result quantifies the variance component for survival forest models. Furthermore, we show with two specific examples how these concentration bounds, combined with properly designed splitting rules, yield consistency results. The development of these results serves as a general framework for showing the consistency of tree- and forest-based survival models.

✉ cuiy@live.unc.edu

ADDITIVE RATES MODEL FOR RECURRENT EVENT DATA WITH INFREQUENTLY OBSERVED TIME-DEPENDENT COVARIATES

Tianmeng Lyu*, *University of Minnesota*

Yifei Sun, *Columbia University*

Chiung-Yu Huang, *University of California, San Francisco*

Xianghua Luo, *University of Minnesota*

Various regression methods have been proposed for analyzing recurrent event data. Among them, the semiparametric additive rates model is appealing because the regression coefficients quantify the absolute difference in the occurrence rate of recurrent events between different groups. Theoretically, the additive rates model permits time-dependent covariates, but model estimation requires the values of time-dependent covariates being observed throughout the follow-up period. In practice, however, time-dependent covariates are usually infrequently observed. In this paper, we propose to kernel smooth functionals of time-dependent covariates across subjects in the estimating function. The proposed method is flexible enough to handle situations where both time-dependent and time-independent covariates are present. It can also accommodate multiple time-dependent covariates, each observed at a different time schedule. Simulation studies show that the proposed method outperforms the LCCF method or the linear interpolation method. The proposed method is illustrated by analyzing data from a study which evaluated the effect of streptococcal infections on recurrent pharyngitis episodes.

✉ lyuxx025@umn.edu

80. CANCER APPLICATIONS

METHODS FOR INTEGRATING METHYLATION AND EXPRESSION DATA FOR PROBING SMOKING EXPOSURE EFFECTS IN MUSCLE-INVASIVE UROTHELIAL CARCINOMA PATIENTS

Miranda L. Lynch*, *Roswell Park Cancer Institute*

Jessica M. Clement, *UConn Health*

Cancer is an inherently genetic disease, and the complex etiology of cancer has led to the development of multiple measurement modalities to probe the disease at fine degrees of molecular detail across multiple levels of information. These include transcriptomic profiles geared towards understanding gene expression, mutational burden, SNPs, and copy number alterations that characterize the disease state. Also, epigenetic analyses give information about the gene regulatory framework and chromatin modifications that impact expression without altering the genetic sequence. Developing methods for integrating this information is key to comprehending the extreme complexity and heterogeneity of cancer progression. In this work, we focus on novel consensus clustering approaches and inferential procedures derived in that framework to specifically probe the epigenetic alterations/gene expression interface as it is differential between smoking status groups in bladder cancer. We illustrate the proposed methodology using Illumina Infinium HumanMethylation 450 BeadChip arrays and RNA-seq based expression data from TCGA bladder cancer patients.

✉ Miranda.Lynch@roswellpark.org

STATISTICAL PROPERTIES OF THE D-METRIC FOR MEASURING ETIOLOGIC HETEROGENEITY IN CASE-CONTROL STUDIES

Emily C. Zabor*, *Memorial Sloan Kettering Cancer Center*

Venkatraman E. Seshan, *Memorial Sloan Kettering Cancer Center*

Shuang Wang, *Columbia University*

Colin B. Begg, *Memorial Sloan Kettering Cancer Center*

As molecular and genomic profiling of tumors has become increasingly common, the focus of cancer epidemiologic research has shifted away from the study of risk factors for disease as a single entity, and toward the identification of risk factors for subtypes of disease. The idea that risk factors for disease may differ across subtypes is known as etiologic heterogeneity. We have previously proposed an approach to the study of etiologic heterogeneity in the context of case-control studies, which integrates dimension reduction of potentially high dimensional tumor marker data through k-means clustering with a search for the most heterogeneous disease subtypes according to the available risk factor data, based on optimizing a scalar measure D. Here we investigate the statistical properties of this approach using simulation studies, and address questions related to how both the number and strength of structure of tumor markers impact the method's ability to accurately identify the etiologically distinct subtypes and approximate the true extent of etiologic heterogeneity.

✉ zabore@mskcc.org

PATHWAY-GUIDED INTEGRATIVE ANALYSIS OF HIGH THROUGHPUT GENOMIC DATASETS TO IMPROVE CANCER SUBTYPE IDENTIFICATION

Dongjun Chung*, *Medical University of South Carolina*

Zequn Sun, *Medical University of South Carolina*

Andrew Lawson, *Medical University of South Carolina*

Brian Neelon, *Medical University of South Carolina*

Linda Kelemen, *Medical University of South Carolina*

Identification of cancer patient subgroups using high throughput genomic data is of critical importance to clinicians and scientists because it can offer opportunities for more personalized treatment and overlapping treatments of cancers. However, it still remains challenging to implement robust and interpretable identification of cancer subtypes and driver molecular features using these massive, complex, and heterogeneous datasets. In this presentation, I will discuss our novel Bayesian framework to identify cancer subtypes and driver molecular features by integrating multiple types of cancer genomics datasets with biological pathway information. I will discuss the proposed method with simulation studies and its application to TCGA datasets.

✉ chungd@musc.edu

A FAST SCORE TEST FOR GENERALIZED MIXTURE MODELS

Rui Duan* •, *University of Pennsylvania*

Yang Ning, *Cornell University*

Shuang Wang, *Columbia University*

Bruce G. Lindsay, *The Pennsylvania State University*

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

Yong Chen, *University of Pennsylvania*

In biomedical studies, testing for homogeneity between two groups, where one group is modeled by mixture models, is often of great interest. This paper considers the semiparametric exponential family mixture model proposed by Hong et al. 2017, and studies the score test for homogeneity under this model. The score test is nonregular in the sense that nuisance parameters disappear under the null hypothesis. To address this difficulty, we propose a modification of the score test, so that the resulting test enjoys the Wilks phenomenon. In finite samples, we show that with fixed nuisance parameters the score test is locally most powerful. In large samples, we establish the asymptotic power func-

tions under two types of local alternative hypotheses. Our simulation studies illustrate that the proposed score test is powerful and computationally fast. We apply the proposed score test to an UK ovarian cancer DNA methylation data for identification of differentially methylated CpG sites.

✉ ruiduan@upenn.edu

STATISTICAL APPROACHES FOR META-ANALYSIS OF GENETIC MUTATION PREVALENCE

Margaux L. Hujoel*, *Harvard School of Public Health/ Dana-Farber Cancer Institute*

Giovanni Parmigiani, *Harvard School of Public Health/ Dana-Farber Cancer Institute*

Danielle Braun, *Harvard School of Public Health/ Dana-Farber Cancer Institute*

Estimating the prevalence of rare genetic mutations in the general population is of great interest as it can inform genetic counseling and risk management. Most studies which estimate prevalence of mutations are performed in high-risk populations, and each study is designed with differing inclusion-exclusion (i.e. ascertainment) criteria. Combining estimates from multiple studies through a meta-analysis is challenging due to the differing study designs and ascertainment mechanisms. We propose a general approach for conducting a meta-analysis under these complex settings by incorporating study-specific ascertainment mechanisms into a likelihood function. We implement the proposed likelihood based approach using both frequentist and Bayesian methodology. We evaluate these approaches in simulations and show that the proposed methods result in unbiased estimates of the prevalence even with rare mutations (a prevalence of 0.01%). An advantage of the Bayesian approach is uncertainty in ascertainment probabilities can be easily incorporated. We apply our methods in an illustrative example to estimate the prevalence of PALB2 in the general population by combining multiple studies.

✉ hujoel@g.harvard.edu

MATHEMATICAL MODELING IDENTIFIES OPTIMUM LAPATINIB DOSING SCHEDULES FOR THE TREATMENT OF GLIOBLASTOMA PATIENTS

Shayna R. Stein*, *Harvard School of Public Health and Dana-Farber Cancer Institute*

Franziska Michor, *Harvard School of Public Health, Dana-Farber Cancer Institute and Harvard University*

Hiroshi Haeno, *Kyushu University, Japan*

Igor Vivanco, *The Institute of Cancer Research, London*

Human primary glioblastomas (GBM) often harbor mutations within the epidermal growth factor receptor (EGFR). Treatment of EGFR-mutant GBM cell lines with the EGFR inhibitor lapatinib can induce cell death. However, EGFR inhibitors have shown little efficacy in the clinic, partly due to inappropriate dosing. Here, we developed a computational approach to model in vitro cell dynamics of the EGFR-mutant cell line SF268 in response to different lapatinib concentrations and dosing schedules. We used this approach to identify an effective treatment within clinical toxicity limits, and developed a partial differential equation model to study in vivo GBM treatment response by taking into account the heterogeneous and diffusive nature GBM. Our model predicts that continuous dosing remains the best strategy for lowering tumor burden compared to pulsatile schedules. Our mathematical modeling and statistical analysis provides a rational method for comparing treatment schedules in search for optimal dosing strategies for GBM.

✉ sstein@g.harvard.edu

8I. PRESIDENTIAL INVITED ADDRESS

STATISTICS AS PREDICTION:

Roderick J. Little, *Professor of Biostatistics, Richard D. Remington Distinguished University Professor, Department of Statistics Research Professor, Institute for Social Research Senior Fellow, Michigan Society of Fellows, University of Michigan*

I have always thought that a simple and unified approach to problems in statistics is from the prediction perspective – the objective is to predict the things you don't know, with appropriate measures of uncertainty. My inferential philosophy is "calibrated Bayes" – Bayesian predictive inference for a statistical model that is developed to have good frequentist properties. I discuss this viewpoint for a number of problems in missing data and causal inference, contrasting it with other approaches.

✉ rlittle@umich.edu

82. POSTERS

CHALLENGES IN DEVELOPING LEARNING ALGORITHMS TO PERSONALIZE TREATMENT IN REAL TIME

Susan A. Murphy*, *Harvard University*

A formidable challenge in designing sequential treatments is to determine when and in which context it is best to deliver treatments. Consider treatment for individuals struggling with chronic health conditions. Operationally designing the sequential treatments involves the construction of decision rules that input current context of an individual and output a recommended treatment. That is, the treatment is adapted to the individual's context; the context may include current health status, current level of social support and current level of adherence for example. Data sets on individuals with records of time-varying context and treatment delivery can be used to inform the construction of the decision rules. There is much interest in personalizing the decision rules, particularly in real time as the individual experiences sequences of treatment. Here we discuss our work in designing online "bandit" learning algorithms for use in personalizing mobile health interventions.

✉ samurphy@fas.harvard.edu

DECISION MAKING TO OPTIMIZE COMPOSITE OUTCOMES

Daniel J. Luckett*, *University of North Carolina, Chapel Hill*

Eric B. Laber, *North Carolina State University*

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

Precision medicine, the idea of tailoring treatment based on individual characteristics, has potential to improve patient outcomes. Individualized treatment rules formalize precision medicine as maps from the covariate space into the treatment space. One statistical task in precision medicine is the estimation of an optimal individualized treatment rule. In many applications, there are multiple outcomes and clinical practice involves making decisions to balance trade-offs. In this setting, the underlying goal is that of optimizing a composite outcome constructed from a utility function of the outcomes. This precludes direct application of existing methods for estimating treatment rules, as the true underlying utility function may be unknown. We propose a method for estimating treatment rules in the presence of multiple outcomes by modeling the decisions made in observational data, estimating the true utility function, and estimating a treatment rule to optimize the resulting composite outcome. We show consistency of the utility function estimator. We demonstrate the performance of the proposed method in simulation and through an analysis of a bipolar disorder study.

✉ luckett@live.unc.edu

A SEQUENTIAL CONDITIONAL TEST FOR MEDICAL DECISION MAKING

Min Qian*, *Columbia University*

Due to patient heterogeneity in response to various aspects of any treatment program, biomedical and clinical research has shifted from the traditional one-size-fits-all treatment to personalized medicine. An important step in this direction is to identify the treatment and covariate interactions. We consider the setting in which there are a potentially large

number of covariates of interest. Although a bunch of novel variable selection methodologies are being developed to aid in treatment selection in this setting, few, if any, has adopted formal hypothesis testing procedures. In this talk, I will present a bootstrap based testing procedure which can be used to sequentially identify variables that interact with treatment. The method is shown to be effective in controlling type I error rate with a satisfactory power as compared to competing methods.

✉ mq2158@cumc.columbia.edu

MODELING SURVIVAL DISTRIBUTION AS A FUNCTION OF TIME TO TREATMENT DISCONTINUATION: A DYNAMIC TREATMENT REGIME APPROACH

Shu Yang*, *North Carolina State University*

Anastasios Tsiatis, *North Carolina State University*

Michael Blazing, *Duke University Medical Center*

We estimate how the treatment effect on the survival distribution depends on the time to discontinuation of treatment. There are two major challenges. First, the formulation of treatment regime in terms of time to treatment discontinuation is subtle. A naive approach is to define the treatment regime "stay on the treatment until time t", which however is not sensible in practice. Our innovation is to cast the treatment regime as a dynamic regime "stay on the treatment until time t or until a treatment-terminating event occurs". Secondly, the major challenge in estimation and inference arises from biases associated with the nonrandom assignment of treatment regimes, because, naturally, treatment discontinuation is left to the patient and the physician and so time to discontinuation depends on the patient's disease status. To address this issue, we develop dynamic-regime Marginal Structural Models and inverse probability of treatment weighting to estimate the impact of time to treatment discontinuation on a survival outcome, compared to the effect of not discontinuing treatment.

✉ syang24@ncsu.edu

83. ADVANCED WEIGHTING METHODS FOR OBSERVATIONAL STUDIES

THE AVERAGE TREATMENT EFFECT ON THE EVENLY MATCHABLE UNITS (ATM): A VALUABLE ESTIMAND IN CAUSAL INFERENCE

Lauren R. Samuels*, *Vanderbilt University School of Medicine*

Robert A. Greevy, *Vanderbilt University School of Medicine*

While the average treatment effect on the treated (ATT) may be of interest in observational studies, many studies that attempt to estimate the ATT are in fact providing either biased estimates of the ATT or possibly unbiased estimates of another quantity altogether. In this presentation we examine this other commonly estimated quantity, which we call the average treatment effect on the evenly matchable units (ATM). We formally define "evenly matchable units" and show that the ATM is estimated by 1:1 matching with a propensity score caliper and by the "matching weights" introduced by Li and Greene (2013). We present three new weighting-based methods for ATM estimation, including Bagged One-to-One Matching (BOOM) weights. By explicitly choosing to use ATM weighting, analysts can focus their inference on the units for whom the least amount of model extrapolation is required.

✉ lauren.samuels@vanderbilt.edu

MATCHING WEIGHTS TO SIMULTANEOUSLY COMPARE THREE TREATMENT GROUPS: COMPARISON TO THREE-WAY MATCHING

Kazuki Yoshida*, *Harvard School of Public Health*

Sonia Hernandez-Diaz, *Harvard School of Public Health*

Daniel H. Solomon, *Brigham and Women's Hospital*

John W. Jackson, *Johns Hopkins Bloomberg School of Public Health*

Joshua J. Gagne, *Brigham and Women's Hospital*

Robert J. Glynn, *Brigham and Women's Hospital*

Jessica M. Franklin, *Brigham and Women's Hospital*

Matching weights (MW) are an extension of IPTW that weights both exposed and unexposed groups to emulate propensity score matching (PSM). We generalized MW to multiple groups and compared the performance in the three-group setting to 1:1:1 PSM and IPTW. We also applied these methods to an empirical example of three analgesics. MW had similar bias, but better MSE compared to three-way matching in all scenarios. The benefits were more pronounced in scenarios with a rare outcome, unequally sized treatment groups, or poor covariate overlap. IPTW's performance was highly dependent on covariate overlap. In the empirical example, MW achieved the best balance for 24 out of 35 covariates. Hazard ratios were numerically similar to PSM. However, the confidence intervals were narrower for MW. MW demonstrated improved performance over 1:1:1 PSM in terms of MSE, particularly in simulation scenarios where finding matched subjects was difficult. Given its natural extension to settings with even more than three groups, we recommend matching weights for comparing outcomes across multiple treatment groups, particularly in settings with rare outcomes or unequal exposure distributions.

✉ kazukiyoshida@mail.harvard.edu

A TALE OF TWO TAILS: ADDRESSING EXTREME PROPENSITY SCORES VIA THE OVERLAP WEIGHTS

Laine E. Thomas*, *Duke University*

Fan Li, *Duke University*

Fan Li, *Duke University*

Inverse probability weighting in causal inference is often hampered by extreme (close to 0 or 1) propensity scores, leading to biased estimates and excessive variance. A common remedy is to trim or truncate extreme propensity scores. However, such methods are often sensitive to cutoff points, and correspond to an ambiguous target population. Overlap weights are a newly developed alternative, in which each unit's weight is proportional to the probability of being assigned to the opposite group. The weights are bounded and minimize the variance of the weighted average treatment effect among the class of balancing weights. By continuously down-weighting the units in the tails of the propensity score distribution, arbitrary trimming of the propensity scores is avoided and the target population emphasizes patients with the most overlap in observed characteristics between treatments. Through analytical derivations and simulations, we will illustrate the advantages of the overlap weights over the standard IPW with trimming or truncation, in terms of reducing bias and variance induced by extreme propensity scores. Joint work with Fan Li (sq).

✉ laine.thomas@duke.edu

EXPLORING FINITE-SAMPLE BIAS IN PROPENSITY SCORE WEIGHTS

Lucy D'Agostino McGowan*, *Vanderbilt University*

Robert Greevy, *Vanderbilt University*

The principle limitation of all observational studies is the potential for unmeasured confounding. Various study designs may perform similarly in controlling for bias due to measured confounders while differing in their sensitivity to unmeasured confounding. Design sensitivity (Rosenbaum, 2004) quantifies the strength of an unmeasured confounder needed to nullify an observed finding. In this presentation, we explore how robust certain study designs are to various unmeasured confounding scenarios. We focus particularly on two exciting new study designs - ATM and ATO weights. We illustrate the performance in a large electronic health records based study and provide recommendations for sensitivity to unmeasured confounding analyses in ATM and ATO weighted studies, focusing primarily on the potential reduction in finite-sample bias.

✉ ld.mcgowan@vanderbilt.edu

84. SPATIAL MODELING OF ENVIRONMENTAL AN EPIDEMIOLOGICAL DATA

BAYESIAN MODELING OF NON-STATIONARY SPATIAL PROCESSES VIA DOMAIN SEGMENTATION

Veronica J. Berrocal*, *University of Michigan*

A key component of statistical models for environmental applications is the spatial covariance function, which is traditionally assumed to belong to a parametric class of stationary models whose parameters are estimated using the observed data. While convenient, the assumption of stationarity is often non-realistic. In this talk we present two Bayesian statistical approaches to model non-stationary environmental processes by assuming that the process is locally stationary. Regions of stationarity are determined differently in the two modeling frameworks: in the first, they are defined as segments of the geographic space were spatially-varying covariates are more homogeneous, in the second they are regions where the spatially-varying scale of the environmental process is more homogeneous. In the first modeling approach, we use Bayesian Model Averaging to account for uncertainty in the segmentation of the geographic space, in the second we express the spatial process using an M-RA basis expansion (Katzfuss 2017) with mixture priors on the basis coefficients. We illustrate the two methodologies with an application in soil science and air pollution.

✉ berrocal@umich.edu

BAYESIAN MODELS FOR HIGH-DIMENSIONAL NON-GAUSSIAN DEPENDENT DATA

Jonathan R. Bradley*, *Florida State University*

A Bayesian approach is introduced for analyzing high-dimensional dependent data that are distributed according to a member from the exponential family of distributions. This problem requires extensive methodological advancements, as jointly modeling high-dimensional dependent data leads to the so-called 'big n problem'. The computational complexity of this problem is further exacerbated by allowing for non-Gaussian data models. Thus, we develop new computationally efficient distribution theory for this setting. In particular, we introduce a class of conjugate multivariate distributions for the exponential family. We discuss several theoretical results regarding conditional distributions, an asymptotic relationship with the multivariate normal distribution, parameter models, and full-conditional distributions for a Gibbs sampler. We demonstrate the modeling framework through several examples, including an analysis of a large environmental dataset.

✉ bradley@stat.fsu.edu

USING POINT PATTERNS TO IDENTIFY PRINCIPAL DRIVERS OF HEAT-RELATED MORBIDITY

Matthew J. Heaton*, *Brigham Young University*

Jacob W. Mortensen, *Simon Fraser University*

Olga V. Wilhelmi, *National Center for Atmospheric Research*

Cassandra Olenick, *National Center for Atmospheric Research*

Persistent, extreme heat is associated with various negative public health outcomes such as heat stroke, heat exhaustion, kidney failure, circulatory and nervous system complications and, in some cases, death. Interestingly, however, these negative public health outcomes are largely preventable using simple intervention strategies such as cooling centers or water fountains. In order to be effective, however, such intervention strategies need to be strategically located. Hence, epidemiologists often construct risk maps pinpointing trouble spots throughout a city so as to identify locations of highest need. In this research, we construct such risk maps from a point pattern of negative public health outcomes in Houston, TX. Specifically, we define a log-Gaussian Cox process model for heat-related morbidity and merge these outcomes via Bayesian hierarchical modeling.

✉ mheaton@stat.byu.edu

MULTIVARIATE SPATIO-TEMPORAL (MVST) MIXTURE MODELING OF HEALTH RISK WITH ENVIRONMENTAL STRESSORS

Andrew B. Lawson*, *Medical University of South Carolina*

Rachel Carroll, *National Institute of Environmental Health Sciences, National Institutes of Health*

It is often the case that researchers wish to simultaneously explore the behavior of and estimate overall risk for multiple, related diseases with varying rarity while accounting for potential spatial and/or temporal correlation. In this presentation, we propose a flexible class of multivariate spatio-temporal mixture models to fill this role. Further, these models offer flexibility with the potential for model selection as well as the ability to accommodate lifestyle, socio-economic, and physical environmental variables with spatial, temporal, or both structures. Here, we explore the capability of this approach via a large scale simulation study and examine a motivating data example involving three cancers in South Carolina. The results which are focused on four model variants suggest that all models possess the ability to recover simulation ground truth and display improved model fit over two baseline Knorr-Held spatio-temporal interaction model variants in a real data application.

✉ lawsonab@musc.edu

85. LATEST DEVELOPMENT OF STATISTICAL METHODS FOR TUMOR HETEROGENEITY AND DECONVOLUTION

ROBUST SUBCLONAL ARCHITECTURE RECONSTRUCTION FROM ~2,700 CANCER GENOMES

Wenyi Wang*, *University of Texas MD Anderson Cancer Center*

Kaixian Yu, *University of Texas MD Anderson Cancer Center*

Hongtu Zhu, *University of Texas MD Anderson Cancer Center*

The composition of subpopulations of cancer cells may affect cancer prognosis and treatment efficacy. Understanding the subclonal structure helps infer the evolution of tumor cells which can further guide the discovery of driver mutations. In the Pan-Cancer Analysis of Whole Genomes (PCAWG) initiative, we were faced with the challenge of characterizing subclonality in an unprecedented set of 2,778 tumor samples from 40 histologically distinct cancer type. Over the course of 4 years, we have encountered and addressed two bottleneck analytical issues, first to accurately call subclonal mutations in the whole-genome sequencing data from paired tumor-normal samples, and then to accurately cluster mutations into clonal and subclonal categories and estimate the corresponding fraction of cells containing these mutations. This talk will recount our effort in developing new statistical methods and tools, including a somatic mutation caller MuSE, a fast subclonal reconstruction caller CliP, and finally a consensus mutation clustering method CSR, for the analysis and biological interpretation of all PCAWG cancer genomes.

✉ wwang7@mdanderson.org

ESTIMATION OF INTRA-TUMOR HETEROGENEITY AND ASSESSING ITS IMPACT ON SURVIVAL TIME

Wei Sun*, *Fred Hutchinson Cancer Research Center*

Chong Jin, *University of North Carolina, Chapel Hill*

Paul Little, *University of North Carolina, Chapel Hill*

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

Mengjie Chen, *University of Chicago*

A tumor sample of a single patient often includes a conglomerate of heterogeneous cells. Understanding such intra-tumor heterogeneity may help us better characterize the tumor sample and identify useful biomarkers to guide the practice of precision medicine. We have developed a new statistical method, SHARE (Statistical method for

Heterogeneity using Allele-specific REads and somatic point mutations), which reconstructs clonal evolution history using whole exome sequencing data of matched tumor and normal samples. Our method jointly models copy number aberrations and somatic point mutations using both total and allele-specific read counts. We further study the association between intra-tumor heterogeneity and survival time, while accounting for the uncertainty of estimating intra-tumor heterogeneity.

✉ wsun@fredhutch.org

CANCER GENOMICS WITH BULK AND SINGLE CELL SEQUENCING

Nancy R. Zhang*, *University of Pennsylvania*

Yuchao Jiang, *University of North Carolina, Chapel Hill*

Zilu Zhou, *University of Pennsylvania*

Cancer is a disease driven by rounds of Darwinian selection on somatic genetic mutations, and recent advances in sequencing technologies is offering new opportunities as well as revealing new challenges in understanding the genetics of cancer. In this talk, I will describe the use of bulk and single cell sequencing to infer a tumor's clonal evolutionary history. First, I will describe a framework that we developed to estimate the underlying evolutionary tree by joint modeling of single nucleotide mutation and allele-specific copy number profiles from repeated bulk sequencing data. Then, I will describe how single cell RNA sequencing data can be harnessed to improve subclone detection and phylogeny reconstruction.

✉ nzh@wharton.upenn.edu

UNDERSTANDING CANCER PROGRESSION VIA TUMOR EVOLUTION MODELS

Russell Schwartz*, *Carnegie Mellon University*

Progression of cancers has long been understood to be an evolutionary phenomenon, although our understanding of that process has been greatly revised as it has become possible to probe genetic variation within and between tumors in ever finer detail. Tumor phylogenetics, which arose to bring methods from evolutionary biology to the interpretation of cancer genomics, has become a key tool for making sense of the complexity of modern genetic variation data sets. In this talk, we examine the utility of tumor evolutionary models for predicting outcomes of cancer progression, such as metastasis or mortality. Such work proceeds from the recognition that heterogeneity in mutation processes between patients, inferred by tumor phylogenies, carries predictive power for future progression beyond what is available from more traditional profiles of specific mutations at a given instance in time. We demonstrate this strategy and consider tradeoffs between distinct technologies and study designs. We close by considering emerging directions in tumor phylogenetics and the challenges new technologies are bringing to phylogeny inference and robust prediction of tumor progression.

✉ russells@andrew.cmu.edu

86. STATISTICAL ANALYSIS OF MICROBIOME DATA

VARIABLE SELECTION FOR HIGH DIMENSIONAL COMPOSITIONAL DATA WITH APPLICATION IN METAGENOMICS

Hongmei Jiang*, *Northwestern University*

Metagenomics is a powerful tool to study the microbial organisms living in various environments. The abundance of a microorganism or a taxon is usually estimated using relative proportion or percentage in sequencing-based metagenomics studies. Due to the constraint of the sum of the relative abundances being 1 or 100%, standard conventional statistical methods may not be suitable for metagenomics data analysis. In this talk we will discuss characterization of the association between microbiome and disease status and variable selection in regression analysis with compositional covariates. We compare the performance of different methods through simulation studies and real data analysis.

✉ hongmei@northwestern.edu

JOINT MODELING AND ANALYSIS OF MICROBIOME WITH OTHER OMICS DATA

Michael C. Wu*, *Fred Hutchinson Cancer Research Center*

Understanding the relationship between microbiome and other omics data types is important both for obtaining a more comprehensive view of biological systems as well as for elucidating mechanisms underlying outcomes and response to exposures. However, such analyses are challenging. Issues inherent to microbiome data include dimensionality, compositionality, sparsity, phylogenetic constraints, and complexity of relationships among taxa. It remains unclear how to address these issues, much less to address these issues in combination with problems specific to other omics data types and problems in modeling relationships between microbial taxa and other omics features. To move towards joint analysis, we propose development of methods for studying both community level correlations between microbiome and other data types as well as for correlating individual taxa with other omics data. Real data analyses demonstrate that our approach for correlating microbial taxa with other omics features can reveal new biological findings.

✉ mcwu@fhcrc.org

A TWO-STAGE MICROBIAL ASSOCIATION MAPPING FRAMEWORK WITH ADVANCED FDR CONTROLLING PROCEDURES

Huilin Li*, *New York University*

Jiyuan Hu, *New York University*

Hyunwook Koh, *New York University*

Linchen He, *New York University*

Menghan Liu, *New York University*

Martin J. Blaser, *New York University*

One special feature of microbiome data is the taxonomical tree which characterizes the microbial evolutionary relationship. Microbes that are taxonomically close usually behave similarly or have similar biological functions. Incorporating and utilizing this microbial dependence structure, we propose a two-stage microbial association testing framework to gain extra power in the microbial taxa discovery. Comparing to the conventional microbial association test which performs the microbial taxa association scan at the target rank taxon by taxon and control the FDR by the BH procedure afterwards, the proposed framework achieve the more powerful result with less multiple comparison penalty. Extensive simulations and real data validation are used to illustrate the superiority of the proposed method.

✉ huilin.li@nyumc.org

A NOVEL APPROACH ON DIFFERENTIAL ABUNDANCE ANALYSIS FOR MATCHED METAGENOMIC SAMPLES

Lingling An*, *University of Arizona*

Wenchi Lu, *University of Arizona*

Di Ran, *University of Arizona*

Dan Luo, *University of Arizona*

Qianwen Luo, *University of Arizona*

Dailu Chen, *University of Texas Southwestern Medical Center*

Many diseases such as cancer, diabetes, and bowel disease are highly associated with human microbiota. Next-generation sequencing technology allows us to detect features/species contained in human microbial communities. Oftentimes, the counts of features are observed as over-dispersed and non-negative count data with excess zeros. Such data lead some differential abundance analysis methods to apply Zero-Inflated Negative Binomial (ZINB) regression for modeling the microbial abundance. In addition, in order to account for the within-subject variation of repeated measurements from the same subject, random effect terms, which are commonly assumed to be independent, are added to the models. In this research, we propose a two-part model cZINB model with correlated random

effects considered, for testing the association between two groups of repeated measurements collected at different conditions for the same subject. Through comprehensive simulation studies, we demonstrate that cZINB outperforms the existing methods in detecting the significantly differential abundant features for matched microbial samples.

✉ anling@email.arizona.edu

87. RECENT ADVANCES IN STATISTICAL METHODS FOR IMAGING GENETICS

MOMENT-MATCHING METHODS FOR HIGH-DIMENSIONAL HERITABILITY AND GENETIC CORRELATION ANALYSIS

Tian Ge*, *Harvard Medical School*

Chia-Yen Chen, *Harvard Medical School*

Mert R. Sabuncu, *Cornell University*

Jordan W. Smoller, *Harvard Medical School*

Heritability and genetic correlation analyses provide important information about the genetic basis of complex traits. With the exponential progress in genomic technologies and the emergence of large-scale data collection efforts, classical methods for heritability and genetic correlation estimation can be difficult to apply when analyzing high-dimensional neuroimaging features or data sets with large sample sizes. We develop unified and computationally efficient (co)heritability analysis methods based on method of moments. We apply our methods to (1) conduct the first comprehensive heritability analysis across the phenotypic spectrum in the UK Biobank and identify phenotypes whose heritability is moderated by age, sex and socioeconomic status; and (2) investigate the shared genetic influences between vertex-wise morphological measurements (e.g., cortical thickness and surface area) derived from structural brain MRI scans, and fluid intelligence and major psychiatric disorders.

✉ tge1@mgh.harvard.edu

FROM ASSOCIATION TO CAUSATION: CASUAL INFERENCE IN IMAGING-GENETIC DATA ANALYSIS

Momiao Xiong*, *University of Texas Health Science Center at Houston*

Nan Lin, *University of Texas Health Science Center at Houston*

Zixin Hu, *Fudan University*

Rong Jiao, *University of Texas Health Science Center at Houston*

Vince D. Calhoun, *The Mind Research Network*

We develop novel structural causal models coupled with integer programming as a new framework for inferring large-scale causal networks of genomic-brain images. The proposed method for large-scale genomic-imaging causal network analysis was applied to the MIND clinical imaging consortium's schizophrenia image-genetic study with 142 series of diffusion tensor MRI images and 50,725 genes typed in 64 schizophrenia patients and 78 healthy controls. Images were segmented into 23 regions. A region was taken as a node. The sparse SEMs were used to compute score of image node. IP was used to search the optimal causal graph. Linear SEMs with IP identified 5 image regions causing SCZ. The ANM narrowed down 5 regions to 3 causal regions: Frontal_R, Occipital_R , and Occipital & Pareital_ Sup. We identified 176 SNPs that were associated with imaging signal variation, 82 SNPs significantly associated with SCZ and 27 SNPs that cause imaging signal variation.

✉ momiao.xiong@uth.tmc.edu

IMAGING-WIDE ASSOCIATION STUDY: INTEGRATING IMAGING ENDOPHENOTYPES IN GWAS

Wei Pan*, *University of Minnesota*

Zhiyuan Xu, *University of Minnesota*

Chong Wu, *University of Minnesota*

A new and powerful approach, called imaging-wide association study (IWAS), is proposed to integrate imaging endophenotypes with GWAS to boost statistical power and enhance biological interpretation for GWAS discoveries. IWAS extends the promising transcriptome-wide association study (TWAS) from using gene expression endophenotypes to using imaging and other endophenotypes with a much wider range of possible applications. As illustration, we use gray-matter volumes of several brain regions of interest (ROIs) drawn from the ADNI-1 structural MRI data as imaging endophenotypes, which are then applied to the individual-level GWAS data of ADNI-GO/2 and a large meta-analyzed GWAS summary statistics dataset (based on about 74000 individuals), uncovering some novel genes significantly associated with Alzheimer's disease (AD). We also compare the performance of IWAS with TWAS, showing much larger numbers of significant AD-associated genes discovered by IWAS, presumably due to the stronger link between brain atrophy and AD than that between gene expression of normal individuals and the risk for AD.

✉ weip@biostat.umn.edu

FUNCTIONAL GENOME-WIDE ASSOCIATION ANALYSIS OF IMAGING AND GENETIC DATA

Hongtu Zhu*, *University of Texas MD Anderson Cancer Center*

The aim of this paper is to develop a functional genome-wide association analysis (FGWAS) framework to efficiently carry out whole-genome analyses of functional phenotypes. FGWAS consists of three components: a multivariate varying coefficient model, a global sure independence screening procedure, and a test procedure. Compared with the standard multivariate regression model, the multivariate varying coefficient model explicitly models the functional features of functional phenotypes through the integration of smooth coefficient functions and functional principal component analysis. Statistically, compared with existing methods for genome-wide association studies (GWAS), FGWAS can substantially boost the detection power for discovering important genetic variants influencing brain structure and function.

✉ hzhu5@mdanderson

88. CLINICAL TRIALS AND BIOPHARMACEUTIAL RESEARCH

FORMULATION OF CONFIDENCE INTERVALS FOR DIFFERENCE BETWEEN TWO BINOMIAL PROPORTIONS FROM LOGISTIC REGRESSION

Ryuji Uozumi*, *Kyoto University School of Medicine*

Shinjo Yada, *A2 Healthcare Corporation*

Kazushi Maruo, *University of Tsukuba*

Atsushi Kawaguchi, *Saga University*

In randomized parallel-group clinical trials, the influence of potentially relevant factors on the outcome must be considered. The use of logistic regression is required for binary data. A common method of reporting the result of logistic regression is to provide an odds ratio and its corresponding confidence interval. However, there is currently no useful method to obtain the confidence interval for the difference between binomial proportions based on logistic regression using available statistical analysis software. Hence, the results of such statistical analyses cannot be further evaluated with respect to the consistency of confidence intervals between the odds ratio and the difference between proportions. In this work, we propose a novel method to construct the confidence intervals for the difference between two binomial proportions based on parameter estimates of logistic regression. The performance of the proposed method is investigated via a simulation study that includes the situation in which the sample size is not large, the proportion is close to 0, and the sample size allocation is unequal. The results from the simulation study will be presented at conference.

✉ uozumi@kuhp.kyoto-u.ac.jp

BIG DATA VS DATA RE-USE: EXAMPLE OF PATIENTS' RECRUITMENT MODELING

Nicolas J. Savy*, *Toulouse Mathematics Institute*

Nathan Minois, *INSERM Unit 1027*

Valerie Lauwers-Cances, *CHU Toulouse*

Stephanie M. Savy, *INSERM Unit 1027*

Michel Attal, *CHU Toulouse*

Sandrine Andrieu, *INSERM Unit 1027; Philippe Saint-Pierre, Toulouse Mathematics Institute*

Vladimir V. Anisimov, *University of Glasgow*

Big Data strategies based on huge databases (data farms) involves marvelous opportunities in Medical Research especially for raising research hypotheses. But for hypothesis testing or for modeling issues, two problems emerges: the large sample size makes estimations meaningless and heterogeneity of data makes predictions inefficient. For dealing with such questions, to re-use the data of existing database (for example early stage of development for drug development) may be an alternative strategy of paramount interest. Indeed, statistical inference is more efficient and models (Bayesian for instance) calibrated from the existing database works pretty well. As an example, the re-use of the recruitment data of a completed clinical trial for dealing with the feasibility of a new clinical trial in the same therapeutic aera and involving more or less the same centres as the completed one (in terms of recruitment) is presented. The methodology is presented and the performance of the model assessed.

✉ Nicolas.Savy@math.univ-toulouse.fr

TOWARD MORE FLEXIBLE AND EFFECTIVE CONTROL OF FALSE POSITIVES IN PHASE III RANDOMIZED CLINICAL TRIALS

Changyu Shen*, *Beth Israel Deaconess Medical Center, Harvard Medical School*

Phase III randomized clinical trials to demonstrate efficacy of a medical intervention is an essential framework that shapes the landscape of current medicine development covering essentially all treatment guidelines and healthcare policies. For this reason, it involves perhaps the most rigorous study design, analysis of the data and interpretation of the results. Uncertainty on the treatment efficacy through statistical evaluation has become a primary factor driving the conclusion of a trial. However, the universally applied threshold of $p<0.05$ may not be appropriate for all trials. In addition, there is a lack of knowledge on the global magnitude of false positives generated by the large number of randomized trials in the United States. We have developed a strategy that controls for the expected number of false positives, which also allows for more flexible type I error control for each individual trial. Central to the strategy is the parameter of the proportion of null or negative efficacy measures among all trial efficacy measures, which can be estimated through results of trials registered at clinicaltrials.gov.

✉ cshen1@bidmc.harvard.edu

TARGETED MAXIMUM LIKELIHOOD ESTIMATION TO IMPROVE PRECISION AND REDUCE BIAS IN ALZHEIMER'S DISEASE CLINICAL TRIALS

Elizabeth Colantuoni*, *Johns Hopkins Bloomberg School of Public Health*

Aidan McDermott, *Johns Hopkins Bloomberg School of Public Health*

Jon Steingrimsson, *Brown University School of Public Health*

Arnold Baker, *Johns Hopkins University School of Medicine*

Michela Gallagher, *Johns Hopkins University School of Medicine*

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

Consider a two arm regulatory trial where the primary outcome is measured at baseline and several fixed follow-up times. The primary endpoint is the change in the primary

outcome from baseline to the final follow-up and the average treatment effect is the difference in mean change comparing the treatment and control arm. Assume that a set of prognostic baseline variables are collected and patient drop-out is expected. To estimate the average treatment effect, the mixed model for repeated measures (MMRM) is the standard approach. However, novel targeted minimum loss estimators (TMLE) proposed by Van der Laan and Gruber (2012) can be applied to this setting and may offer reductions in bias and precision gains in relative to MMRM. We use data from the Alzheimer's Disease Cooperative Study to simulate hypothetical clinical trials for a drug that reduces the decline in cognitive impairment among persons with mild cognitive impairment. We compare key statistical properties of MMRM and TMLE for estimating the average treatment effect when varying the prognostic ability of the baseline variables and the models generating patient drop-out.

✉ elizabethcolantuoni@gmail.com

CLINICAL TRIAL SIMULATION AND VIRTUAL PATIENTS' GENERATION

Philippe Saint Pierre*, *Toulouse Mathematics Institute*

Nicolas J. Savy, *Toulouse Mathematics Institute*

Clinical trial simulations (CTS) have been recognized by the pharmaceutical companies and regulatory authorities as being pivotal to improving the efficiency of the drug development process. This includes the use of CTS to learn about drug effectiveness and safety and to optimize trial designs at the various stages of development. A CTS is the study of the effects of a drug in virtual patient populations using mathematical models. The generation of virtual patients is a key point of the process. Three approaches are considered for this task. The Discrete method is based on Monte Carlo simulations from the joined distribution of the covariates whereas the Continuous method used a multi-normal distribution. A third approach using R-vines copula is proposed for patients generation. Modelling patients'evolution over time is another challenging task. As an example, data on 30000 HIV patients are used to evaluate various scenario of (approaching) switching to generics of many VIH drugs. Execution models are developed to update patients characteristics (including treatment) over time. Simulations are performed to evaluate the cost of various switching to generics scenarios.

✉ Philippe.Saint-Pierre@math.univ-toulouse.fr

89. CLUSTERED AND HIERARCHICAL DATA

THE MIXTURE APPROACH TO ESTIMATING A POPULATION-AVERAGED VALUE

Haoyu Zhang*, *Johns Hopkins Bloomberg School of Public Health*

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

When analyzing hierarchical data with the target of inference a population-level feature computed from the unit-specific features, the analyst needs to take control of the estimation process. For example, if the target is an equally weighted average of unit-specific values, the single-parameter MLE will be minimum variance but may incur a large bias penalty. If unit-specific estimates are unbiased, their mean is also unbiased, but the approach may incur a substantial variance penalty. And, if the unit-specific estimates are themselves biased or not available for all units, an alternative approach is needed. We consider the use of mixture modeling to estimate equally weighted or unequally weighted functionals of unit-specific parameters, with focus on evaluating performance of the Non-parametric Maximum Likelihood (NPML) mixture model. We compare performance to parametric mixture modeling and use of a single-parameter likelihood for the geometric sampling distribution. We illustrate the methods using data from the VenUS I trial.

✉ hzhang71@jhu.edu

A CLASS OF PRIOR DISTRIBUTIONS FOR ADJACENCY MATRICES IN CONDITIONAL AUTOREGRESSIVE MODELS

Heli Gao*, *Florida State University*

Jonathan Bradley, *Florida State University*

Traditional conditionally autoregressive (CAR) models use neighborhood information to define the adjacency matrix. Specifically, the neighborhoods are formed deterministically using the boundaries between the regions. However, areas far apart may be highly correlated, and CAR models are unable to identify this situation. We propose a class of prior distributions for adjacency matrices, which can detect a relationship between two areas that do not share a boundary. Our approach is fully Bayesian, and involves a computationally efficient conjugate update of the adjacency matrix. To illustrate the high performance of our Bayesian hierarchical model, we present a simulated study, and an example using the environmental dataset.

✉ gaoheli0303@gmail.com

A THEORETICAL EXPLORATION OF CLUSTERABILITY TESTS

Naomi C. Brownstein*, *Florida State University*

Clusterability is a newer topic related to clustering that measures the inherent cluster structure in a dataset. Applied before clustering, a clusterability test serves as a pre-clustering validation procedure. The goal of a clusterability test is to select one of two outcomes. First, a clusterability test tells the user whether or not the data was generated from a multimodal distribution, which would correspond to multiple clusters. Second, the tests aim to stop clustering users from clustering data generated from a single unimodal distribution, where clustering would generally be inappropriate. Due to the widespread use of clustering throughout biometrics, especially genetics, there is a need for wider and deeper understanding of clusterability tests. Clusterability tests have been shown to perform well on simulated and real data, but their theoretical properties have not been studied. This talk explores theoretical properties of clusterability methods for data generated from one or more unimodal continuous distributions.

✉ naomi.brownstein@med.fsu.edu

A STOCHASTIC SECOND-ORDER GENERALIZED ESTIMATING EQUATIONS APPROACH FOR ESTIMATING INTRACLASS CORRELATION IN THE PRESENCE OF INFORMATIVE MISSING DATA

Tom Chen*, *Harvard School of Public Health*

Eric Tchetgen Tchetgen, *Harvard School of Public Health*

Rui Wang, *Harvard School of Public Health*

Design and analysis of cluster randomized trials must take into account correlation among outcomes from the same clusters. When applying standard generalized estimating equations (GEE), the first-order (e.g. treatment) effects can be estimated consistently even with a misspecified correlation structure. In settings for which the correlation is of interest, one could estimate this quantity via second-order generalized estimating equations (GEE2). We build upon GEE2 in the setting of missing data, for which we incorporate a "second-order" inverse-probability weighting (IPW) scheme and "second-order" doubly robust (DR) estimating equations that guard against partial model misspecification. We highlight the need to model correlation among missingness indicators in such settings. In addition, the computational difficulties in solving these second-order equations have motivated our development of more computationally efficient algorithms for solving GEE2, which alleviates reliance on parameter starting values and provides substantially faster and higher convergence rates than the more widely used deterministic root-solving methods.

✉ tomchen@g.harvard.edu

BAYESIAN HIERARCHICAL MODEL IN PHASE I DOSE ESCALATION STUDY WITH DIFFERENT ETHNIC GROUPS

Serena Liao*, *Novartis Oncology Pharmaceuticals*

In oncology drug development, populations from different ethnic groups may show slightly different safety profiles and separate dose escalation processes (e.g. a Western population and an Asian population) may be operated. In a typical Bayesian logistic regression model, the historical data or data from other populations can be used to derive the prior through meta-analytic-predictive (MAP) approach. Posterior distribution of DLT rate will be updated with new data. However, in the situation where the dose escalations for different ethnic groups are running in parallel, the MAP prior has to be re-derived at each dose escalation for data borrowing. The meta-analytic-combined (MAC) approach applied in this work is equivalent to MAP approach, yet allowing the incorporation of evolving data in an adaptive way. Both co-data and historical data can be used in the Bayesian hierarchical model we built, where exchangeability parameters are defined. The proposed methodology induces adaptive borrowing based on the similarity of the data. Performance of the proposed model and adaptive information borrowing mechanism across populations will be explored by a real life example.

✉ liaoge.serena@gmail.com

BAYESIAN HIERARCHICAL MODELLING OF AIR POLLUTION EXTREMES USING MULTIVARIATE MAX-STABLE PROCESSES

Sabrina Vettori* •, *King Abdullah University of Science and Technology*

Raphael Huser, *King Abdullah University of Science and Technology*

Marc Genton, *King Abdullah University of Science and Technology*

Capturing the potentially strong dependence between peak exposures of multiple air pollutants across a large spatial region is crucial for assessing the associated public health risks.

In order to investigate the multivariate spatial dependence properties of air pollution extremes, we introduce a new class of multivariate max-stable processes. Our proposed model admits a hierarchical formulation, in which the data are conditionally independent given some latent nested τ-stable random factors, facilitating Bayesian inference and offering a convenient and interpretable tree-based characterisation. We fit this nested multivariate max-stable model to air pollution concentration and temperature maxima collected at a number of sites in the Los Angeles area, showing that the proposed model succeeds in capturing their complex tail dependence structure.

✉ sabrina.vettori@kaust.edu.sa

90. EXPERIMENT DESIGN

DESIGN ANALYSES OF RANDOMIZED CLINICAL TRIALS SUPPORTING FDA CANCER DRUG APPROVALS

Emily Lord, *Boston University School of Public Health*

Isabelle R. Weir, *Boston University School of Public Health*

Ludovic Trinquart*, *Boston University School of Public Health*

Background: The conventional design and interpretation of randomized controlled trials (RCT) emphasizes statistical significance. Methods: We reviewed pivotal RCTs supporting FDA approval of cancer drugs between 2007-2016. We performed design analyses of the RCTs for overall survival (OS) and progression-free survival (PFS). We estimated the type S error risk - concluding that the new treatment is beneficial when it is actually detrimental - and the exaggeration ratio - the factor by which the magnitude of the estimated treatment effect differs from the true effect. Results: We analyzed

43 trials for 39 approved drugs. For a true HR of 0.7 on OS, the median type S error risk was 0.00% [Q1-Q3, 0.00%-0.01%] and the exaggeration ratio 1.09 [1.01-1.11]. These numbers were 3.56% [0.40%-6.74%] and 1.30 [1.13-1.42] for a true HR of 0.9. The latter suggests a 30% overestimation of the true effect. Results were similar for PFS. Conclusion: Our findings highlight the value of design analysis for RCTs and offer a quantification of the winner's curse, in which pivotal RCTs supporting cancer drug approval tend to be biased and overly optimistic.

✉ ludovic@bu.edu

SAMPLE SIZE CONSIDERATIONS FOR COMPARING DYNAMIC TREATMENT REGIMENS IN A SEQUENTIAL MULTIPLE-ASSIGNMENT RANDOMIZED TRIAL WITH A CONTINUOUS LONGITUDINAL OUTCOME

Nicholas J. Seewald*, *University of Michigan*

Kelley M. Kidwell, *University of Michigan*

James R. McKay, *University of Pennsylvania*

Inbal Nahum-Shani, *University of Michigan*

Daniel Almirall, *University of Michigan*

Clinicians and researchers alike are increasingly interested in how best to individualize interventions. A dynamic treatment regimen (DTR) is a sequence of pre-specified decision rules which guides the delivery of an individualized sequence of treatments that is tailored to specific and possibly changing needs of the individual. The sequential multiple-assignment randomized trial (SMART) is a research tool which allows for the construction of effective DTRs. We introduce a method for computing sample size for SMARTs in which the primary aim is to compare two embedded DTRs using a continuous repeated-measures outcome collected over the entire study. The sample size method is based on a longitudinal analysis that accounts for unique features of a SMART design. These features include modeling constraints and the over- or under-representation of different sequences of treatment (by design). We illustrate our methods using the ENGAGE study, a SMART aimed at developing a DTR for increasing motivation to attend treatment sessions among alcohol- and cocaine-dependent patients.

✉ nseewald@umich.edu

OPTIMAL DESIGN OF REPLICATION EXPERIMENTS

Ryan T. Jarrett*, *Vanderbilt University*

Matthew S. Shotwell, *Vanderbilt University*

Many scientific fields are increasingly concerned that much of their published experimental findings cannot be replicated. While the discussion around this issue has garnered many useful suggestions, little has been said about how replication experiments should be conducted. To address this issue, we introduce a general approach to efficiently designing replication studies. We use the tools of optimal design and the information provided in the original study to identify sampling points that provide the most information about the outcome. The extent of replication is measured by the intersection between the confidence region on the difference in parameter estimates, and a pre-specified indifference region about the null. Consequently, the criteria that we minimize are functions of the joint covariance matrix for the difference in parameter estimates. In simulations, the results of this method are shown to out-perform those of random sampling under several models and are largely robust to poor parameter estimates in the original studies. Extensions to criteria used in robust optimal design approaches and to sequential optimal design approaches are additionally considered.

✉ ryan.t.jarrett@vanderbilt.edu

DESIGN OF NONINFERIORITY RANDOMIZED TRIALS USING THE DIFFERENCE IN RESTRICTED MEAN SURVIVAL TIMES

Isabelle R. Weir*, *Boston University School of Public Health*

Ludovic Trinquart, *Boston University School of Public Health*

The design of noninferiority trials is challenging, often requiring very large sample sizes. For time-to-event outcomes, differences in Restricted Mean Survival Times (RMSTD), an alternative to Hazard Ratios (HR), could lead to smaller trial sizes. We redesigned 35 noninferiority trials and compared the required sample sizes based on the two measures. We tested for non-proportional hazards (NPH) and for noninferiority. We calculated the RMST margin equivalent to the HR margin and found the required sample size for both, using alpha and beta from the original design. We found evidence of NPH in 15% of trials. The two measures had consistent conclusions for noninferiority except in one trial. The median HR margin was 1.49. The median of the RMSTD margins was -21 days for a median time horizon of 2 years. When using the RMSTD instead of the HR, the required sample size was smaller in 71% of trials (median relative decrease, 8.5%, Q1-Q3 0.4%-38.0%), sparing 25000 participants from enrollment. The HR margins may seem large but can translate to trivial RMSTD margins. The RMSTD can result in considerable reductions in required sample size.

✉ iweir@bu.edu

9I. NONLINEAR AND SEMI-PARAMETRIC MODELS

IMPROVED BAYESIAN SCHEME FOR RESOLVING INTRAVOXEL NEUROANATOMY

Sharang Chaudhry*, *University of Nevada, Las Vegas*

Kaushik Ghosh, *University of Nevada, Las Vegas*

The brain, for the purpose of medical imaging, is often viewed as a collection of cubic elements called voxels. To delineate nerve pathways in the brain, it is typically required to understand the neuroanatomy locally at the voxel level. This local estimation problem is nonlinear and underdetermined with large number of nuisance parameters. We propose a method using the Reversible Jump Markov Chain Monte Carlo (RJMCMC) strategy on an additive tensor signal model to resolve the number and direction of nerves within voxels. The use of RJMCMC sampler is intuitive since we perform model estimation and discrimination simultaneously. Furthermore, we use a set of constrained conditional priors to mitigate issues of identifiability. Finally, we explore the performance of the method on simulated and clinical datasets, and perform comparative analyses with existing methods.

✉ sharang.chaudhry@unlv.edu

IMPROVING ESTIMATION OF GENERALIZED SEMI-PARAMETRIC VARYING-COEFFICIENT MODELS USING COVARIANCE FUNCTION

Fang Fang*, *University of North Carolina, Charlotte*

Yanqing Sun, *University of North Carolina, Charlotte*

This paper studies a generalized semi-parametric varying-coefficient model for longitudinal data that can flexibly model three types of covariate effects: time-constant effects, time-varying effects, and covariate-varying effects. We investigate a profile weighted least square approach for model estimation by utilizing within subject correlations. Several methods for incorporating the within subject correlations are explored including quasi-likelihood approach(QL), minimum generalized variance approach (MGV),the quadratic inference function approach (QIF) and newly proposed minimum weighted residuals approach. The asymptotic properties of the estimators are derived theoretically. Our simulation study shows that the covariance assisted estimation is more efficient then working independence approach. The proposed estimation methods are applied to a real data set.

✉ fangfangxmu704@gmail.com

SEMIPARAMETRIC COMPARISON OF NONLINEAR CURVES AND SURFACES

Shi Zhao*, *Indiana University School of Medicine*

Spencer George Lourens, *Indiana University School of Medicine*

Giorgos Bakoyannis, *Indiana University School of Medicine*

Wanzhu Tu, *Indiana University School of Medicine*

Despite the increased popularity of semiparametric estimation of curves and surfaces in clinical investigations, comparisons of nonlinear functions have not been well studied. Few existing methods are available as off-the-shelf tools for practical data analysis. Existing methods also have important limitations, such as lack of accommodation of covariates and repeatedly measurements, and thus further restricting their potentials for application. Herein, we propose a wild bootstrap procedure within the penalized semiparametric regression framework. The method can be used for comparisons of curves and surfaces with both cross-sectional and longitudinal data. Compared to existing methods, the proposed testing procedure has an added flexibility of accommodating linear covariates. Preliminary simulation suggests a satisfactory performance in terms of Type I error rate control and analytical power. The method is implemented in an R package with a user-friendly interface for general use. We illustrate the use of the method in a real clinical example.

✉ zhaoshi169@gmail.com

TESTING NONLINEAR GENE-ENVIRONMENT INTERACTION THROUGH VARYING COEFFICIENT AND LINEAR MIXED MODELS

Zhengyang Zhou*, *Southern Methodist University and University of Texas Southwestern Medical Center*

Hung-Chih Ku, *DePaul University*

Chao Xing, *University of Texas Southwestern Medical Center*

We present a novel statistical procedure to detect the nonlinear gene-environment (GxE) interaction with continuous traits in sequencing association studies. Commonly-used approaches for GxE interaction usually assume linear relationship between genetic and environmental factor, thus they suffer power loss when the underlying relationship is nonlinear. A varying-coefficient model (Ma et al., 2011) is proposed to relax the linear assumption, however, it's unable to adjust for population stratification, a major source of confounding in genome-wide association studies. To overcome these limitations, we develop the Varying-Coefficient embedded Linear Mixed Model (VC-LMM) for assessing the nonlinear GxE interaction and accounting for population stratification. The proposed VC-LMM well controls type I error rates when the population stratification is present, and it's powerful for both common and rare variants. We apply computationally efficient algorithms for generating null distributions and estimating parameters in the linear mixed model, thus the computational burden is greatly reduced. Using simulation studies, we demonstrate the performance of VC-LMM.

✉ zhengyangz@smu.edu

CHANGE POINT INFERENCE IN PRESENCE OF MISSING COVARIATE

Tao Yang*, *Fred Hutchinson Cancer Research Center*

Ying Huang, *Fred Hutchinson Cancer Research Center*

Threshold regression is plausible to model the relationship between the risk of infectious disease to the vaccine-induced surrogate biomarker response. Despite its appealing interpretation, limited research has been conducted to study the model when covariates are missing, which is the case that in a standard vaccine trial, the vaccine-induced biomarker in placebo group is missing. We focus on studying the linear threshold regression model with interaction term

for missing covariates, and an iterative algorithm will be proposed for parameter estimation. The maximum of the likelihood ratio statistic based on the estimated likelihood across a sequence of threshold/change point will be proposed to test the existence of change point. The asymptotic distribution of the proposed estimator will be studied, and finite sample properties of the estimator including both estimation and inference will be shown in simulation studies.

✉ tyang23@fredhutch.org

CONSTRUCTING CONCURRENT NETWORK OF BIOMARKER PROCESSES USING DYNAMICAL SYSTEMS

Ming Sun*, *Columbia University*

Donglin Zeng, *University of North Carolina, Chapel Hill*

Yuanjia Wang, *Columbia University*

Modeling dynamical system with a large number of components faces the challenge of how to extract structural and regulatory information from noisy observations. In this work, we propose model and method for network construction and parameter estimation using noisy, high-dimensional, multiple time-course data observed from ordinary differential equations (ODEs) systems. Our work is motivated by studies of brain responses to stimuli on patients affected by psychiatric disorders using experiments collecting electroencephalogram (EEG) data. We use single-index model to capture the relationship between the derivatives of each target state variable and all regulatory state variables in the system. To jointly model multiple time-course datasets, we extend our method to double-index model to incorporate subject-level or experimental level features. We propose an integrated index model for ordinary differential equations (IIM-ODE) and a hybrid algorithm for estimation. The network structure sparsity is achieved by introducing regularization in the step of estimating network association parameters.

✉ ms4799@cumc.columbia.edu

MULTIVARIATE SKEWED RESPONSES: NEW SEMIPARAMETRIC REGRESSION MODEL AND A BAYESIAN RECOURSE

Apurva Chandrashekhar Bhingare* •, *Florida State University*

Debajyoti Sinha, *Florida State University*

Debdeep Pati, *Texas A&M University*

Stuart R. Lipsitz, *Harvard Medical School*

Dipankar Bandyopadhyay, *Virginia Commonwealth University*

For many clinical studies with skewed multivariate responses, carefully accommodating the level of skewness and association structure are essential for accurate inference and prediction of the covariate effects on the response of interest. We present a novel semiparametric model with nonparametric error density and associated theoretically justifiable semiparametric Bayesian analysis of such studies. Similar to multivariate Gaussian densities, our multivariate model is closed under marginalization and allows a wide class of multivariate associations. Compared to existing models, our model enjoys several desirable properties, including meaningful physical interpretations of skewness levels and covariate effects on the marginal density, Bayesian computing via available software, ease of prior specifications and assurance of consistent Bayesian estimates of parameters. We illustrate the practical advantages of our methods over existing parametric alternatives via application to a periodontal study and via a simulation study.

✉ a.bhingare@stat.fsu.edu

92. PREDICTION AND PROGNOSTIC MODELING

TUNING PARAMETER SELECTION FOR PREDICTION IN HIGH-DIMENSIONAL RIDGE REGRESSION

Zilin Li*, *Harvard School of Public Health*

Lee Dicker, *Rutgers University*

Xihong Lin, *Harvard School of Public Health*

We consider in this paper tuning parameter selection for prediction in high-dimensional ridge regression. Classical asymptotic predictive results for tuning parameter selection in ridge regression are in fixed design and the existing asymptotic predictive results for random design in ridge regression rely on non-data driven tuning parameter or strong assumption of the covariate distribution. In view of selecting optimal tuning parameter for prediction under general cases, we derive finite sample bound on the predictive loss ratio for tuning parameter selection criteria Cp, GCV, AIC and BIC. Our finite sample results are under mild assumption of the covariate matrix and allow for both fixed and random design. Based on our concentration bounds, we study the asymptotic optimality of these criterions. We performed simulation studies to evaluate the prediction performance of these criteria over a broad range of the covariate matrix structure. Our simulation results coincided with the proposed theoretical results and showed that Cp outperforms other criteria when the number of predictors is less than the number of observations.

✉ li@hsph.harvard.edu

A BAYESIAN METHOD FOR UPDATING WEIBULL PREDICTION MODELS USING PUBLISHED SUMMARY DATA

Wen Ye*, *University of Michigan*

Pin Li, *University of Michigan*

Clinical prediction models (CPM) are useful tools for providing risk estimates for individual patients to guide medical decisions. There is often a need to update existing prediction models for new settings in time and place. Existing methods for updating CPM requires using individual level data or combining a set of published CPMs. For updating Weibull prediction models, we discuss a variety of Bayesian algorithms for integrating summary data provided in different formats in the literature, including marginal survival Kaplan-Meier curves and reported cumulative counts for the same outcome. Reported summary statistics on baseline patient characteristics are used to account for heterogeneity within and between studies. Using data from multiple sources, this method allows updating not only the scale and shape parameters but also regression coefficients in a Weibull prediction model. This method is evaluated via a comprehensive simulation study. In addition, we apply our method to update the UKPDS (United Kingdom Prospective Diabetes Study) Outcome I stroke model using data from recently published clinical studies.

✉ wye@umich.edu

DISCRIMINATION INDEX FOR MULTI-CATEGORY OUTCOME

Aya Kuchiba*, *National Cancer Center, Japan*

Kentaro Sakamaki, *University of Tokyo*

Prediction models for multi-category outcome would provide the risk vectors containing the estimated absolute risks for each category. Compared with binary outcome, assessing discrimination performance for multi-category outcome is not straightforward, because the "concordance" can be defined in many ways. The volume under the receiver operating characteristic (ROC) surface (VUS) has been proposed

for three-category outcomes, which is the probability that the outcome of a randomly selected set of subjects is all correctly identified. VUS has been extended to the outcome with more than three categories as hypervolume under the ROC manifold (HUM). Alternative measure is the polytomous discrimination index (PDI), which can be interpreted as the probability that the subject in the set of subjects with distinct true outcome status is correctly classified into one's own category. We introduce a new measure of discrimination index for multi-category outcome, which is based on the concordance of each component of risk vectors from the set of subjects with distinct true outcome status. We also characterize the behavior of proposed index with comparing to PDI and HUM.

✉ akuchiba@ncc.go.jp

EVALUATING DISCRIMINATORY ACCURACY OF MODELS USING PARTIAL RISK-SCORES IN TWO-PHASE STUDIES

Parichoy Pal Choudhury*, *Johns Hopkins University*

Anil K. Chaturvedi, *National Cancer Institute, National Institutes of Health*

Nilanjan Chatterjee, *Johns Hopkins University*

Before clinical applications, risk prediction models need to be evaluated in independent studies (e.g., prospective cohort studies) that did not contribute to model development. Often prospective cohort studies ascertain information on some expensive biomarkers in a nested sub-study of the original cohort, typically selected based on case-control status and additional covariates. We propose an efficient approach for evaluating Area Under the Curve (AUC) using data from all individuals irrespective of whether they were sampled in the sub-study. The approach involves estimating probabilities of risk-scores for cases being larger than those in controls conditional on partial risk-scores as opposed to multivariate risk factor profiles. This allows estimation of the underlying conditional probabilities using subjects with complete covariate information in a non-parametric fashion even when numerous covariates are involved. We evaluate finite sample performance of the proposed method and compare it to an inverse probability weighted (IPW) estimator through extensive simulation studies. We apply the method to evaluate a lung cancer risk prediction model using data from the PLCO trial.

✉ parichoy@jhu.edu

MACHINE LEARNING ALGORITHMS FOR SURVIVAL, LONGITUDINAL, AND MULTIVARIATE (SLAM) DATA WITH APPLICATIONS TO SUDDEN CARDIAC ARREST (SCA)

Shannon Wongvibulsin*, *Johns Hopkins University School of Medicine*

Katherine Wu, *Johns Hopkins University School of Medicine*

Scott Zeger, *Johns Hopkins Bloomberg School of Public Health*

Studies of sudden cardiac arrest (SCA), a leading cause of death, produce repeated measures on risk factors and non-SCA outcomes that impact SCA risk. We develop Random Forest for SLAM (RF-SLAM) data to predict SCA risk and determine the relative contribution of each risk factor to an individual's overall SCA risk. Our RF-SLAM method partitions the multiple events and predictors for each individual into a set of what we term "Counting Process Information Units" (CPIUs). Each CPIU contains the person indicator, interval indicator, multivariate outcome values, summary function values of outcome history, predictor values, and length of the interval. RF-SLAM uses the general random forest approach to estimate the multivariate hazard within each unit of time. We develop methods for assembling and visualizing the time varying hazards and survival functions for each individual. With data from the Prospective Observational Study of Implantable Cardioverter-Defibrillators (PROSE-ICD), we illustrate the use of RF-SLAM for determining population risk as well as for predicting individualized SCA risk to guide treatment decisions.

✉ swongvi1@jhmi.edu

ZERO-INFLATED QUANTILE REGRESSION WITH ITS APPLICATION IN NOMAS

Wodan Ling*, *Columbia University*

Bin Cheng, *Columbia University*

Ying Wei, *Columbia University*

Ying Kuen Cheung, *Columbia University*

The Northern Manhattan Study (NOMAS) is a population-based study designed to evaluate the impact of medical, socio-economic, and other risk factors on the incidence of vascular disease in a multi-ethnic, stroke-free cohort, which consists of 3,298 participants recruited between 1993 and 2001. In NOMAS, building a reliable prediction model for plaque burden will lead to an improved prediction of stroke related outcomes in a distant future. Carotid plaque measurements for total plaque area and density were available via high-resolution B-mode ultrasounds in NOMAS. These plaque phenotypes (area and density) take on non-negative values, with a point mass at 0, which corresponds to the group of subjects who could not be detected with a plaque. In this application, we propose a new quantile regression model for zero-inflated, non-negative data. The new model is compared to direct linear quantile regression in simulation studies and NOMAS applications, and shows advantages in producing more accurate estimations and predictions.

✉ wl2459@columbia.edu

INCORPORATING INTER-STUDY HETEROGENEITY INTO THE TRAINING OF PREDICTORS

Prasad Patil*, *Harvard School of Public Health/ Dana-Farber Cancer Institute*

Giovanni Parmigiani, *Harvard School of Public Health/ Dana-Farber Cancer Institute*

We explore a setting where multiple sets of patient data are available to train a predictor. These datasets may exhibit inter-study heterogeneity due to geography, selection criteria, or data processing decisions. As a result, the same feature and outcome may exhibit different associations within each study and yield different decision rules if all studies are considered separately. Oftentimes, these single-study decision rules do not replicate well externally. Options for training predictors in the multi-study setting include merging all datasets together and ignoring heterogeneity; directly accounting for heterogeneity (e.g. meta-analysis); indirectly accounting for heterogeneity by ensembling predictors. We examine the use of weighted ensembling using the cross-study performance of each predictor as well as traditional methods such as unweighted ensembling and stacking. We describe the characteristics of these different options in simulation and in a real-data setting with fifteen ovarian cancer gene expression datasets. We show that each option can work well depending on the amount of inter-study heterogeneity and choice of learning algorithm.

✉ ppatil@jimmy.harvard.edu

93. STATISTICAL GENETICS AND GENOMICS

STATISTICAL APPROACHES TO DECREASING THE DISCREPANCY OF NON-DETECTS IN qPCR DATA

Valeriia Sherina* •, *University of Rochester Medical Center*

Helene McMurray, *University of Rochester Medical Center*

Tanzy M. Love, University of Rochester Medical Center

Matthew N. McCall, *University of Rochester Medical Center*

Quantitative PCR (qPCR) is a widely used methods to measure gene expression. Important problem of qPCR is the presence of non-detects – reactions failing to produce a minimum amount of signal. Most current software replaces non-detects with a value representing the limit of detection,

but this introduces substantial bias in estimation of gene expression. We propose to treat non-detects as non-random missing data, model the missing data mechanism, and use this model to impute missing values or obtain direct estimates of relevant model parameters. To account for the uncertainty inherent in the imputation, we propose a multiple imputation procedure. Three sources of variability are incorporated in introduced methods: parameter estimates, missing data mechanism, and measurement error. We demonstrated the applicability of these methods on real qPCR data, and performed an extensive simulation study to assess model sensitivity. Developed methods are implemented in the R/Bioconductor package nondetects. Introduced statistical methods reduce discrepancies in gene expression, providing more confidence in generating scientific hypotheses and performing downstream analysis.

✉ valeriia_sherina@urmc.rochester.edu

DIFFERENTIAL SPLICING ANALYSIS USING A COMPOSITIONAL REGRESSION METHOD

Scott Van Buren*, *University of North Carolina, Chapel Hill*

Naim Rashid, *University of North Carolina, Chapel Hill*

Changes in relative isoform usage between conditions, known as differential splicing (DS), are often of scientific interest because these changes can cause significant functional differences. Existing approaches for DS analysis often have speed and scalability issues as the number of samples increases. We propose overcoming these issues by using a method designed to analyze compositional data. Specifically, we transform a multivariate vector of relative abundance proportions for isoforms of a specific gene using the isometric log ratio transform. This transforms the data from an n-dimensional vector in a simplex to an (n-1)-dimensional vector on the real line, thereby enabling the use of multivariate ANOVA analysis to examine DS using the condition as a predictor. This framework is flexible in accounting for potential confounding factors and across testing situations, and scales well as the number of samples increases. Using human RNA-seq data, our compositional method results in a greater than 60-fold speed improvement over DRIMSeq, and a simulation study based on the same data finds that the power between the two methods is comparable across effect sizes.

✉ skvanburen@gmail.com

STATISTICAL METHODS FOR PROFILING 3-DIMENSIONAL CHROMATIN INTERACTIONS FROM REPETITIVE REGIONS OF GENOMES

Ye Zheng*, *University of Wisconsin, Madison*

Ferhat Ay, *La Jolla Institute for Allergy and Immunology*

Sunduz Keles, *University of Wisconsin, Madison*

Recent developments in chromatin conformation capture-based assays enabled the high throughput study of 3D chromosomal architecture. In particular, Hi-C elucidated genome-wide long-range interactions among loci. Although the number of statistical methods for Hi-C data is growing rapidly, a key impediment is their inability to accommodate reads aligning to multiple locations thus hinders the comprehensive investigation of interactions involving repetitive regions. We developed mHi-C, a multi-mapping strategy for Hi-C data, as a hierarchical model to probabilistically allocate multi-mapping reads. mHi-C model is built on clustering of sequencing reads that represent biological signals and acknowledge the general features of Hi-C data. Application of mHi-C on published datasets revealed an average increase of 20% in the number of usable reads which translated into higher reproducibility of contact matrices across biological replicates and led to novel significant contacts from heterochromatin regions. Further analysis of newly detected contacts for potential enhancer-promoter interactions highlighted the importance of long-range contacts originating from duplicated segments.

✉ yzheng74@wisc.edu

COMPARISON OF WEIGHTING APPROACHES FOR GENETIC RISK SCORES IN GENE-ENVIRONMENT INTERACTION STUDIES

Anke Huels*, *IUF-Leibniz Research Institute for Environmental Medicine*

Ursula Kraemer, *IUF-Leibniz Research Institute for Environmental Medicine*

Tamara Schikowski, *IUF-Leibniz Research Institute for Environmental Medicine*

Katja Ickstadt, *TU Dortmund University*

Holger Schwender, *Heinrich Heine University*

Weighted genetic risk scores (GRS), defined as weighted sums of risk alleles of single nucleotide polymorphisms (SNPs), are statistically powerful for detecting gene-environment (GxE) interactions. As gold standard, weights are determined externally from independent studies. However, appropriate external weights are not always available. We present two weighting approaches for such situations. The GRS-MI approach uses internal weights from marginal genetic effects, while the GRS-IT approach employs parts of the data to estimate weights from interaction terms with the remaining data being used to determine the GRS. Power and type I error were evaluated in a simulation study for the detection of GxE interactions. In situations with predominant interaction effects, i.e. when SNPs are chosen because of their impact on the biological mechanisms mediating the environmental effect, the highest power was reached with GRS-IT. With predominant marginal genetic effects, GRS-MI was more appropriate. The power of these approaches was only slightly lower than with external weights. In conclusion, weights can be determined from the study population itself, if external weights are unavailable.

✉ anke.huels@iuf-duesseldorf.de

IDENTIFICATION OF CONDITIONALLY ESSENTIAL GENES IN TRANSPOSON SEQUENCING STUDIES

Lili Zhao*, *University of Michigan*

Tn-Seq is a high throughput technique for analysis of transposon mutant libraries to determine conditional essentiality of a gene under an experimental condition. A special feature of the Tn-seq data is that multiple mutants in a gene provides independent evidence to prioritize that gene as being essential. The existing methods either ignore this feature or rely on a high-density transposon library. Moreover, these methods are unable accommodate complex designs. We describe a new, efficient method specifically designed for the analysis of Tn-Seq data. It utilizes two steps to estimate the conditional essentiality for each gene in the genome. First, it collects evidence of conditional essentiality for each insertion by comparing read counts of that insertion between conditions. Second, it combines insertion-level evidence for the corresponding gene. It deals with data from both low- and high-density transposon libraries and accommodates complex designs.

✉ zhaolili@umich.edu

94. SURVIVAL ANALYSIS IN EPIDEMIOLOGY

JOINT MODELING OF RECURRENT AND TERMINAL EVENTS IN NESTED CASE-CONTROL STUDIES

Ina Jazic*, *Harvard School of Public Health*

Sebastien Haneuse, *Harvard School of Public Health*

Virginie Rondeau, *Université de Bordeaux*

The process by which patients experience a series of recurrent events (e.g. hospitalizations, tumor recurrence) may be subject to death, constituting a special case of the semi-competing risks setting. In the complete data setting,

a joint frailty model for the hazards of the recurrent event and death may be used to explore covariate effects on the two event types accounting for their dependence. However, when certain covariates are difficult to obtain, researchers may need to sub-sample patients on whom to collect complete data. One strategy is the nested case-control (NCC) design, in which risk set sampling is performed based on a single outcome. We propose a novel framework for estimation and inference for a joint frailty model for recurrence and death using data from an NCC study, under multiple schemes for risk set formation. We propose a maximum weighted penalized likelihood approach using flexible parametric models for the baseline hazards, investigating operating characteristics as well as design considerations via a simulation study. We illustrate our methods using data from a study on local recurrence, distal metastasis, and death in breast cancer.

✉ ijazic@fas.harvard.edu

MULTIPLICATIVE RATES MODEL FOR RECURRENT EVENTS WITH CASE-COHORT DATA

Poulami Maitra*, *University of North Carolina, Chapel Hill*

Jianwen Cai, *University of North Carolina, Chapel Hill*

Leila D. Amorim, *Federal University of Bahia*

In large prospective cohort studies, accumulation of covariate information and follow-up data make up the majority of the cost involved in the study. This might lead to the study being infeasible when there are some expensive variables and/or the event is rare. Prentice (1986) proposed the case-cohort study for time to event data to tackle this problem. There has been a lot of literature on the application of case-cohort design to univariate and clustered failure time data where the clusters are formed among different individuals. However, recurrent event data are quite common in biomedical and public health research. In this paper, we propose a general case-cohort sampling scheme for recurrent events. We consider multiplicative rates model for recurrent events and proposed an estimating equations approach for parameter estimation. The proposed estimators are shown to be consistent and follow asymptotic normal distribution. The asymptotic approximation works well in finite samples in the simulation studies. We applied the proposed method to data from the Acute Lower Respiratory Tract Infection (ALRI) study among young children in Brazil.

✉ poulamim@live.unc.edu

SEMIPARAMETRIC INFERENCE FOR A TWO-STAGE OUTCOME-DEPENDENT SAMPLING DESIGN WITH INTERVAL-CENSORED FAILURE TIME DATA

Qingning Zhou*, *University of North Carolina, Chapel Hill*

Jianwen Cai, *University of North Carolina, Chapel Hill*

Haibo Zhou, *University of North Carolina, Chapel Hill*

We propose a two-stage outcome-dependent sampling design and inference procedure for studies that concern interval-censored failure time outcomes. This design enhances the study efficiency by allowing the selection probabilities of the second-stage sample, for which the expensive exposure variable is ascertained, to depend on the first-stage observed interval-censored failure time outcomes. In particular, the second-stage sample is enriched by selectively including subjects who are known or observed to experience the failure at an early or late time. We develop a sieve semiparametric maximum estimated likelihood procedure that makes use of all available data from the proposed two-stage design. The resulting regression parameter estimator is shown to be consistent and asymptotically normal, and a consistent estimator for its asymptotic variance is derived. Simulation results demonstrate that the proposed method performs well in practical situations and is more efficient than the simple random sampling design and the adapted inverse probability weighting approach. An application to diabetes data from the Atherosclerosis Risk in Communities (ARIC) study is provided.

✉ qzhou8@uncc.edu

EFFICIENT SECONDARY ANALYSIS OF DATA FROM TWO-PHASE STUDIES

Yinghao Pan*, *University of North Carolina, Chapel Hill*

Jianwen Cai, *University of North Carolina, Chapel Hill*

Elizabeth Jensen, *Wake Forest University*

Haibo Zhou, *University of North Carolina, Chapel Hill*

Under two-phase sampling designs, information on observed event times, event indicator, and easy to obtain covariates is collected in the first phase. Then first phase information such as time to event outcome is used to determine which cohort members will be measured for expensive exposures in the second phase. With tremendous cost involved in collecting the covariates information, it is desirable for investigators to re-use the existing data to study the relationship between covariates and other outcomes of interest. This is referred to as secondary analysis. However, secondary analysis for data from two-phase studies is not easy as the data is not a simple random sample from the general population. This paper provides efficient secondary analysis procedures for data from two-phase studies. The estimation algorithm is based on the maximization of a restricted semiparametric likelihood corresponding to the data structure of the two-phase studies. The resulting estimators are shown to be efficient and asymptotically normal. Data from Norwegian Mother and Child Cohort Study (MoBa) is used to illustrate our method.

✉ ypan@fredhutch.org

TWO-STAGE PSEUDO LIKELIHOOD APPROACH TO ESTIMATION AND INFERENCE FOR RECURRENT EVENTS DATA: APPLICATION TO READMISSION TIME ANALYSIS

Qing Li*, *University of Iowa*

Gideon Zamba, *University of Iowa*

Inpatient hospitalizations account for one-third of the annual healthcare costs in the USA. It is to be noted however that a large percentage of hospital readmissions can be avoided or prevented. At the University of Iowa Hospitals and Clinics, a nurse-led transitional care team (TCT) intervention is deployed in order to prevent unnecessary hospital readmissions. TCT is designed in a way to provide patients with disease self-management, medical education and clear instructions regarding discharge and hospital revisit. In this study we explore the effect of TCT intervention versus a Control, in a quasi-randomization type of analysis based on propensity score matching. By using a two-stage pseudo likelihood approach to estimation and inference for recurrent events data, we analyzed the inter-readmission times and explored the gap-time survival differences between TCT and Control.

✉ qing-li@uiowa.edu

95. NOVEL STATISTICAL APPROACHES FOR ESTIMATING HEALTH EFFECTS OF COMPLEX ENVIRONMENTAL EXPOSURES

DISCOVERING STRUCTURE IN MULTIPLE OUTCOMES MODELS FOR MULTIPLE ENVIRONMENTAL EXPOSURE EFFECTS

Tanzy Love*, *University of Rochester*

Amy LaLonde, *Eli Lilly and Company*

Sally W. Thurston, *University of Rochester*

Phil W. Davidson, *University of Rochester*

Bayesian model-based clustering provides a powerful and flexible tool that can be incorporated into regression models to explore several different questions related to the grouping of observations. In our application, we explore the combined effects of prenatal methylmercury (neurotoxicant) and long-chain-PUFA (neuroprotective) exposure on childhood neurodevelopment. Rather than cluster individual subjects,

we cluster test outcomes within a multiple outcomes model to improve estimation of the exposure effect and the model fit diagnostics. By using information on both exposures in the data to nest the outcomes into groups called domains, the model more accurately reflects the shared characteristics of neurodevelopmental domains. The paradigm allows for sampling from the posterior distribution of the grouping parameters; thus, inference can be made on group membership and their defining characteristics. We avoid the often difficult requirement of a priori identification of the total number of groups by incorporating a Dirichlet process prior. In doing so, we estimate exposure effects on neurodevelopment by shrinking effects within and between the domains selected by the data.

✉ tanzy_love@urmc.rochester.edu

NOVEL TESTS OF MEASUREMENT INVARIANCE IN FACTOR MODELS WITH APPLICATIONS TO ENVIRONMENTAL EPIDEMIOLOGY

Brisa N. Sanchez*, *University of Michigan*

Zhenzhen Zhang, *AbbVie Inc.*

Latent variable models are increasingly used in health areas, including environmental epidemiology. A key assumption in these models is that of measurement invariance: the assumption that the associations between the observed items and latent variables, e.g., factor loadings, are constant for all covariate values. We show examples where this assumption fails, and demonstrate that violating this assumption induces bias in other model parameters, such as exposure-health association estimates. We develop novel tests of measurement invariance (MI) for factor models by modeling factor loadings as varying coefficients. Varying coefficients are estimated using penalized splines, where spline coefficients are treated as random coefficients. MI is tested via a likelihood ratio test for the null hypothesis that the variance of the random spline coefficients equals zero. We use a Monte-Carlo EM algorithm for estimation, and obtain the likelihood using Monte-Carlo integration. Using simulations, we compare the Type I error and power of our testing approach and the multi-group testing method.

✉ brisa@umich.edu

HIERARCHICAL MODELS FOR ESTIMATING ASSOCIATIONS BETWEEN AIR POLLUTION AND HEALTH IN MULTICITY STUDIES

Jenna R. Krall*, *George Mason University*

Howard H. Chang, *Emory University*

Stefanie Ebelt Sarnat, *Emory University*

Multicity studies estimating health effects of short-term air pollution exposure are critical for both understanding city-to-city heterogeneity in health effects and for estimating regional-level health effects associated with pollution exposure. In time series studies using data from many cities, hierarchical models have been extensively applied to estimate health effects associated with air pollution exposure. However, when data are available for only a few cities, qualitative approaches are commonly applied to compare estimated health effects across cities. We developed a Bayesian hierarchical modeling framework to estimate associations between pollution and health in small multicity studies. Our approach facilitates comparisons across a cities by incorporating additional information, such as estimated health effects for several related health outcomes. In five US cities, we estimated associations between 12 air pollutants and emergency department (ED) visits for cardiorespiratory diseases. We examined between-city heterogeneity in estimated health effects to synthesize evidence across exposure-outcome combinations for multiple pollutants and ED visits for specific diagnoses.

✉ jkrall@gmu.edu

BAYESIAN VARYING COEFFICIENT KERNEL MACHINE REGRESSION TO ASSESS COGNITIVE TRAJECTORIES ASSOCIATED WITH EXPOSURE TO COMPLEX MIXTURES

Shelley H. Liu*, *Icahn School of Medicine at Mount Sinai*

Jennifer F. Bobb, *Kaiser Permanente Washington Health Research Institute*

Birgit Claus Henn, *Boston University School of Public Health*

Lourdes Schnaas, *National Institute of Perinatology, Mexico*

Martha M. Tellez-Rojo, *National Institute of Public Health, Mexico*

David Bellinger, *Harvard School of Public Health*

Manish Arora, *Icahn School of Medicine at Mount Sinai*

Robert O. Wright, *Icahn School of Medicine at Mount Sinai*

Brent A. Coull, *Harvard School of Public Health*

Exposure to complex mixtures during early life may exert wide-ranging effects on children's neurodevelopment and cognitive trajectories. However, there is a lack of statistical methods that can accommodate the complex exposure-response relationship between metal mixtures and neurodevelopment, while simultaneously estimating cognitive trajectories. We introduce Bayesian Varying Coefficient Kernel Machine Regression (BVCKMR), a hierarchical model that estimates how mixture exposures at a given time point are associated with neurodevelopmental trajectories. BVCKMR flexibly captures the exposure-response relationship, incorporates prior knowledge, and accounts for non-linear and non-additive effects of individual exposures. Using contour plots and cross-sectional plots, BVCKMR provides information about interaction between complex mixture components. BVCKMR is applied to a subset of data from PROGRESS, a prospective birth cohort study in Mexico City on metal mixture exposures and temporal changes in neurodevelopment.

✉ shelley.liu@mountsinai.org

96. STATISTICAL APPROACHES FOR HANDLING IMPORTANT CHALLENGES FACING CURRENT AGING RESEARCH

WHAT EXACTLY ARE WE MEASURING? HARMONIZATION OF ASSESSMENTS OF OLDER ADULTS

Karen J. Bandeen-Roche*, *Johns Hopkins Bloomberg School of Public Health*

The study of human aging is replete with "constructs" that are multidimensional or cannot at present be measured with single, simple assessments: Disability, frailty, and cognition, for example. Multiple assessments are then utilized to represent the target of measurement: Whether they represent that target comparably across all assessment instances may then be in question. As one example, specific measures may change over repeated instances of measurement; as another, same measures may differentially assess the same health state across subgroups. In this talk, the issues are delineated within a latent variable modeling framework. Methods for addressing them are delineated; implications for the selection of anchoring items and for identifying assessments in need of harmonization are described. The methods are illustrated using data on human aging.

✉ kbandee1@jhu.edu

DESIGN AND DATA FEATURES THAT MAY AFFECT THE ESTIMATION OF THE ONSET OF ACCELERATED COGNITIVE DECLINE

Graciela Muniz Terrera*, *University of Edinburgh*

Eric Peres Barbosa, *Universidad Estatal de Sao Paulo*

Tatiana Benaglia, *Universidad Estatal de Sao Paulo*

Change point models have been used in multiple investigations of cognitive ageing with the purpose of identifying the onset of accelerated decline in cognition. Most commonly used formulations of change point models assume all individuals experience a change point, and either model it as a fixed or, less commonly, as a random effect. Results are mixed and vary by the context of the research question (preclinical dementia, terminal decline, ageing related decline) and the cognitive function evaluated. Yet, features related to the data and study design may also bias change point estimates. In this talk, using simulation studies, we will investigate data features that may bias the estimation of change point models.

✉ g.muniz@ed.ac.uk

INFERRING DIAGNOSTIC ACCURACY FOR CLUSTERED ORDINAL DIAGNOSTIC GROUPS IN THREE-CLASS OR EVEN HIGHER CASE-- APPLICATION TO THE EARLY DIAGNOSIS OF ALZHEIMER'S DISEASE

Chengjie Xiong*, *Washington University in St. Louis*

Jingqin Luo, *Washington University in St. Louis*

Randall Bateman, *Washington University in St. Louis*

Many medical diagnostic studies involve three or more ordinal diagnostic populations in which the diagnostic accuracy can be summarized by the volume or partial volume under the Receiver Operating Characteristic (ROC) surface or hyperplane for a diagnostic marker. When the diagnostic population is clustered, e.g., by families, we propose to model the diagnostic marker by a general linear mixed model that takes into account of the correlation on the diagnostic marker from members of the same clusters. This model then facilitates the maximum likelihood estimation and statistical inferences of the diagnostic accuracy for the diagnostic marker. This approach allows the incorporation of covariates and missing data when some clusters do not have subjects on all diagnostic groups in the estimation of, and the inferences on the diagnostic accuracy. We study the performance of the proposed methods in a simulation study and apply the proposed methodology to the biomarkers database collected by the Dominantly Inherited Alzheimer Network (DIAN).

✉ chengjie@wustl.edu

TRANSLATING ALZHEIMER'S DISEASE RISK POLYMORPHISMS INTO FUNCTIONAL CANDIDATES

Yuriko Katsumata*, *University of Kentucky*

Peter T. Nelson, *University of Kentucky*

Steven Estus, *University of Kentucky*

David W. Fardo, *University of Kentucky*

Alzheimer's disease (AD) is the most common form of dementia. Although the amyloid τ(Aτ) protein and hyperphosphorylated tau aggregates in the brain are considered to be the key pathological hallmarks of AD, the exact cause of AD is yet to be identified. The International Genomics of Alzheimer's Project (IGAP), a consortium that has a goal of characterizing the genetic landscape of AD, revealed significant associations between 19 single nucleotide polymorphisms (SNPs) and AD phenotype. However, most of the pathogenetic loci are located on intronic or intergenic regions, and thus their functional impact are only poorly understood to date. To evaluate the roles of the non-coding SNPs identified in the IGAP, we analyzed whether the non-coding SNPs were either (1) a proxy for an exonic coding variant or (2) associated with altered mRNA transcript levels, by integrating data from many rich resources. For the first hypothesis, rs6656401 in CR1, rs9271192 in HLA-DRB5-DRB1, rs9331896 in CLU, and rs983392 in MS4A6A may be proxies of coding SNPs. For the second hypothesis, rs6656401 in CR1, rs10838725 in CELF1, and rs8093731 in DSG2 acted as an eQTL for several Alzheimer's-associated genes. Our approach for identifying proxies and examining eQTL lessens the impact of the crude gene assignment, although this still remains an open question in the field.

✉ katsumata.yuriko@uky.edu

97. RECENT ADVANCES IN THE ESTIMATION OF GRAPHICAL AND COVARIANCE MODELS

ARMA CHOLESKY FACTOR MODELS FOR THE COVARIANCE MATRIX

Michael J. Daniels*, *University of Florida*

Keunbaik Lee, *Sungkyunkwan University*

Changryong Baek, *Sungkyunkwan University*

In longitudinal studies, serial dependence of repeated outcomes must be taken into account to make correct inferences on covariate effects. As such, care must be taken in modeling the covariance matrix. However, estimation of the covariance matrix is challenging because there are many parameters in the matrix and the estimated covariance matrix should be positive definite. To overcomes these limitations, two Cholesky decomposition approaches have been proposed: modified Cholesky decomposition for autoregressive (AR) structure and moving average Cholesky decomposition for moving average (MA) structure, respectively. Unfortunately, the correlations of repeated outcomes are often not captured parsimoniously using either approach separately. In this paper, we propose a class of flexible, nonstationary, heteroscedastic models that exploits the structure allowed by combining the AR and MA modeling of the covariance matrix that we denote as ARMACD. We analyze a recent lung cancer study to illustrate the power of our proposed methods.

✉ mdaniels@stat.ufl.edu

BAYESIAN HIERARCHICAL MODELING FOR INFERENCE OF MULTIPLE GRAPHICAL MODELS

Christine B. Peterson*, *University of Texas MD Anderson Cancer Center*

Nathan Osborne, *Rice University*

Francesco C. Stingo, *University of Florence*

Marina Vannucci, *Rice University*

I will discuss novel approaches for the joint estimation of multiple graphical models, where the same set of variables are measured across heterogeneous conditions. By taking a Bayesian approach, we are able to formulate flexible hierarchical models which can encourage shared structure or shared edge values, resulting in improved performance for network learning and precision matrix estimation. Methods will be illustrated with an application to understanding changes in brain structural connectivity during the progression of Alzheimer's Disease.

✉ cbpeterson@gmail.com

INFERRING DYNAMIC FUNCTIONAL CONNECTIVITY FROM MAGNETOENCEPHALOGRAPHY RECORDINGS USING TIME-VARYING STATE-SPACE MODELS

Nicholas J. Foti*, *University of Washington*

Adrian K.C. Lee, *University of Washington*

Emily B. Fox, *University of Washington*

Determining the dynamic brain interactions underlying cognitive behaviors is imperative for a variety of problems arising in neuroscience such as understanding the basis for neurological disorders. Magnetoencephalography (MEG) measures the magnetic field produced from neuronal activity and has become an invaluable source of data. Of particular interest is inferring directed interactions between prespecified regions (ROIs). Instead, we use a time-varying linear dynamical system to learn dynamic directed interactions between ROIs directly from MEG sensor recordings. Unlike previous approaches, our state-space model can be applied to many ROIs and multiple subjects. However, efficiently estimating the model parameters is challenging with multiple subjects. This motivates us to develop a stochastic-EM algorithm. We apply the method to MEG recordings of 16 subjects performing auditory attention tasks. The method is able to discern brain interactions known to be associated with auditory attention as well as uncover evidence for hypothesized interactions.

✉ nfoti@uw.edu

COVARIANCE MODELS FOR STRUCTURED SPARSE SHRINKAGE

Peter D. Hoff*, *Duke University*

Maryclare Griffin, *University of Washington*

It is often known that the effects of predictors in high-dimensional regression problems may be related. For example, the effects of predictors corresponding to a particular point in time, point in space, or categorical group are likely to be similar. In such cases, this information can be made use of by shrinking the parameter estimates of similar variables in a similar manner. In this talk I present a model-based approach to inducing structured sparsity in the estimates of related parameters using an adaptive prior covariance model.

✉ peter.hoff@duke.edu

98. STATISTICAL METHODS FOR CANCER RADIOMICS

DEVELOPMENT OF A MULTIPARAMETRIC MR CLASSIFIER FOR PROSTATE CANCER

Joseph Koopmeiners*, *University of Minnesota*

Jin Jin, *University of Minnesota*

Lin Zhang, *University of Minnesota*

Ethan Leng, *University of Minnesota*

Greg Metzger, *University of Minnesota*

Multiparametric magnetic resonance (MR) imaging represents a powerful tool for developing a non-invasive, user-independent tool for the detection of prostate cancer. Previously, our group showed that a voxel-wise classifier that combined quantitative MR parameters from multiple modalities resulted in better classification of prostate cancer than any single parameter, alone. Anatomically, the prostate can be segmented into multiple zones, with the largest being the central gland and peripheral zone, and a primary limitation of our original model is that we ignored the anatomical structure of the prostate when developing our classifier. In this talk, we discuss approaches to accounting for the anatomical structure of the prostate in a voxel-wise multiparametric classifier for prostate cancer. Two approaches will be discussed: a Bayesian approach, which uses the likelihood to model the anatomical structure of the prostate, and an ensemble learning approach that averages local classifiers from sub-regions of the prostate. We will compare the performance of the two classifiers to each other and to a simple model that does not account for the anatomical structure of the prostate.

✉ koopm007@umn.edu

DISTRIBUTION-FREE LIVER CANCER DETECTION USING CT PERFUSION IMAGING

Yuan Wang*, *Washington State University*

CT Perfusion is an emerging non-invasive functional imaging modality that quantifies characteristics pertaining to the passage of fluid through the blood vessel and thus offers a promising quantitative basis for cancer detection, prognostication, and treatment monitoring. In this work, we propose a distribution-free framework for discriminating between regions of liver that contain pathologically verified metastases from healthy liver tissue using the multivariate perfusion parameters. A non-parametric approach is adopted to estimate the joint distribution of the perfusion parameters. Moreover, a Bayesian predictive framework is implemented for simultaneous classification of the spatially correlated regions of interest. The proposed method does not require any distributional assumption and provides more accurate classification by incorporating the spatial correlation between regions.

✉ yuan.wang.stat@gmail.com

A BAYESIAN HIDDEN POTTS MIXTURE MODEL FOR ANALYZING LUNG CANCER PATHOLOGICAL IMAGES

Qiwei Li*, *University of Texas Southwestern Medical Center*

Faliu Yi, *University of Texas Southwestern Medical Center*

Faming Liang, *Purdue University*

Xinglei Wang, *Southern Methodist University*

Yang Xie, *University of Texas Southwestern Medical Center*

Adi Gazdar, *University of Texas Southwestern Medical Center*

Guanghua Xiao, *University of Texas Southwestern Medical Center*

Digital pathology imaging of tumor tissues, which capture histological details in high resolution, is fast becoming a routine clinical procedure. Recent developments in deep-learning have enabled the identification and classification of cells from pathology images at large scale. This creates opportunities to study the spatial patterns of and interactions among different types of cells. We consider the problem of modeling a pathology image with three types of cells: lymphocyte, stromal, and tumor cells. We propose a novel Bayesian hierarchical model, which incorporates a hidden Potts model to project the irregularly distributed cells to a square lattice and an MRF prior model to identify regions in a heterogeneous pathology image. The model allows us to quantify the interactions between different types of cells. We use MCMC sampling techniques, combined with the double Metropolis-Hastings algorithm to sample from the posterior distribution with an intractable normalizing constant. The proposed model was applied to the pathology images of 205 lung cancer patients, and the results show that the interaction strength between tumor and stromal cells predicts patient prognosis.

✉ liqiwei2000@gmail.com

A BAYESIAN NONPARAMETRIC APPROACH FOR CANCER RADIOMICS: ELUCIDATING TEXTURAL PATTERN HETEROGENEITY OF SOLID LESIONS

Xiao Li*, *University of Texas Health Science Center at Houston and University of Texas MD Anderson Cancer Center*

Brian Hobbs, *University of Texas MD Anderson Cancer Center*

Chaan Ng, *University of Texas MD Anderson Cancer Center*

Michele Guindani, *University of California, Irvine*

Cancer radiomics is an emerging tool in cancer diagnostic, due the promise to characterization of lesion phenotypes. A predominate technique for image texture analysis relies on the construction of Gray-Level Co-occurrence Matrices (GLCM). Several approaches have been devised to describe various tissue textural patterns based on sets of feature summary statistics. Reducing the multivariate functional structure inherent to GLCM to sets of summary statistics, current practice is limiting however. In this article, we develop a Bayesian multivariate probabilistic framework for GLCMs wherein we link the observed multivariate counts data to the latent underlying continuous process. The latent spatial association characterizing conditional independence under Gaussian graphs is introduced via a non-parametric Bayesian approach. This approach facilitates to capture the latent cancer subtype leading to a natural clustering of subjects with similar GLCM patterns through sharing of information. Both simulation studies and application to a motivating adrenal lesion imaging dataset reveal the advantages of proposed method over other alternatives.

✉ xiao.li.1@uth.tmc.edu

99. ADVANCING THE ANALYSIS OF MULTIWAY (TENSOR) DATA

SUPERVISED MULTIWAY FACTORIZATION

Gen Li*, *Columbia University*

Eric Lock, *University of Minnesota*

We describe a probabilistic PARAFAC/CANDECOMP (CP) factorization for multiway (i.e., tensor) data that incorporates auxiliary covariates, SupCP. SupCP generalizes the supervised singular value decomposition (SupSVD) for vector-valued observations, to allow for observations that have the form of a matrix or higher-order array. Such data are increasingly encountered in biomedical research and other fields. We use a novel likelihood-based latent variable representation of the CP factorization, in which the latent variables are informed by additional covariates. We give conditions for identifiability, and develop an EM algorithm for simultaneous estimation of all model parameters. SupCP can be used for dimension reduction, capturing latent structures that are more accurate and interpretable due to covariate supervision. Moreover, SupCP specifies a full probability distribution for a multiway data observation with given covariate values, which can be used for predictive modeling. We conduct comprehensive simulations to evaluate the SupCP algorithm, and we apply it to a facial image database with facial descriptors (e.g., smiling / not smiling) as covariates.

✉ gl2521@cumc.columbia.edu

SUPERVISED MODELING OF TENSOR OBJECTS

Rajarshi Guhaniyogi*, *University of California, Santa Cruz*

Shaan Qamar, *Google Inc.*

David B. Dunson, *Duke University*

This talk proposes a Bayesian approach to regression with a tensor predictor or response. Tensor covariates or responses are commonly vectorized prior to analysis, failing to exploit the structure of the tensor, and resulting in poor estimation and predictive performance. We develop a novel class of multiway shrinkage priors for the coefficients in tensor regression models. Properties are described, including posterior consistency under mild conditions, and an efficient Markov chain Monte Carlo algorithm is developed for posterior computation. Simulation studies illustrate substantial gains over vectorizing or using existing tensor regression methods in terms of estimation and parameter inference. The approach is further illustrated in a neuroimaging application.

✉ rajarshign84@gmail.com

STANDARD ERRORS FOR REGRESSION ON RELATIONAL DATA WITH EXCHANGEABLE ERRORS

Bailey K. Fosdick*, *Colorado State University*

Frank W. Marrs, *Colorado State University*

Tyler H. McCormick, *University of Washington*

Relational arrays represent interactions or associations between pairs of actors, often in varied contexts or over time. Such data appear as, for example, financial transactions between individuals, contact frequencies between children, and dynamic protein-protein interactions. This talk proposes and evaluates a new class of parameter standard errors for models that represent elements of a relational array as a linear function of observable covariates. Uncertainty estimates for regression coefficients must account for both heterogeneity across actors and dependence arising from relations involving the same actor. Existing estimators of parameter standard errors that recognize such relational dependence rely on estimating extremely complex, heterogeneous structure across actors. Leveraging an exchangeability assumption, we derive parsimonious standard error estimators that pool information across actors and are substantially more accurate than existing estimators in a variety of settings. We show that our estimator is consistent and demonstrate improvements in inference through simulation and a data set involving international trade.

✉ bailey.fosdick@colostate.edu

COVARIATE-ADJUSTED TENSOR CLASSIFICATION IN HIGH-DIMENSIONS

Xin Zhang*, *Florida State University*

In contemporary scientific research, it is of great interest to predict a categorical response based on a high-dimensional tensor (i.e. multi-dimensional array) and additional covariates. This mixture of different types of data leads to challenges in statistical analysis. Motivated by applications in science and engineering, we propose a comprehensive and interpretable discriminant analysis model, called CATCH model (in short for Covariate-Adjusted Tensor Classification in High-dimensions), which efficiently integrates the covariates and the tensor to predict the categorical outcome. The CATCH model jointly models the relationships among the covariates, the tensor predictor, and the categorical response. More importantly, it preserves and utilizes the intrinsic structure of the data for maximum interpretability and optimal prediction.

✉ henry@stat.fsu.edu

I00. MODERN STATISTICAL METHODS FOR THE EHR ERA

QUANTILE DECISION TREES AND FOREST WITH ITS APPLICATION FOR PREDICTING THE RISK (POST-TRAUMATIC STRESS DISORDER) PTSD AFTER EXPERIENCED AN ACUTE CORONARY SYNDROME

Ying Wei*, *Columbia University*

Classification and regression trees (CART) are a classic statistical learning method that efficiently partitions the sample space into mutually exclusive subspaces with the distinctive means of an outcome of interest. It is a powerful tool for efficient subgroup analysis and allows for complex associations and interactions to achieve high prediction accuracy and stability. Hence, they are appealing tools for precision health applications that deal with large amounts of data from EMRs, genomics, and mobile data and aim to provide a transparent decision mechanism. Although there is a vast literature on decision trees and random forests, most algorithms identify subspaces with distinctive outcome means. The most vulnerable or high-risk groups for certain diseases are often patients with extremely high (or low) biomarker and phenotype values. However, means-based partitioning may not be effective for identifying patients with extreme phenotype values. We propose a new regression tree framework based on quantile regression \cite{Koenker-Bassett1978} that partitions the sample space and predicts the outcome of interest based on conditional quantiles of the outcome variable. We implemented and evaluated the performance of the conditional quantile trees/forests to predict the risk of developing PTSD after experiencing an acute coronary syndrome (ACS), using an observational cohort data from the REactions to Acute Care and Hospitalization (REACH) study\cite{onge2017depressive} at New York Presbyterian Hospital. The results show that the conditional quantile based trees/forest have better discrimination power to identify patients with severe PTSD symptoms, in comparison to the classical mean based CART.

✉ yw2148@cumc.columbia.edu

DISTRIBUTED LEARNING FROM MULTIPLE EHR DATABASES: CONTEXTUAL EMBEDDING MODELS FOR MEDICAL EVENTS

Qi Long*, *University of Pennsylvania*

Ziyi Li, *Emory University*

Xiaoqian Jiang, *University of California, San Diego*

Electronic health records (EHRs) data offer great promises in personalized medicine. However, EHRs data also present analytical challenges due to their irregularity and complexity. In addition, analyzing EHR data involves privacy issues and sharing such data across multiple institutions/sites may be infeasible. A recent work by Farhan et al. (2016) uses contextual embedding models and successfully builds one predictive model for more than seventy common diagnoses.

Although the existing model can achieve a relatively high predictive accuracy, it cannot build global models without sharing data among sites. In this work, we proposed a novel distributed method to learn from multiple databases and build predictive models: Distributed Noise Contrastive Estimation (Distributed NCE). We also extend the proposed method with Differential Privacy to obtain reliable data privacy protections. Our numerical studies demonstrate that the proposed method can build predictive models in a distributed fashion with privacy protection and the resulting models achieve comparable prediction accuracy compared with existing methods that use pooled data across all sites.

✉ qlong@pennmedicine.upenn.edu

LEARNING MATERNAL SMOKING EFFECT ON CHILDHOOD BRONCHIOLITIS FROM TENNCARE

Qingxia Chen*, *Vanderbilt University Medical Center*

David Schlueter, *Vanderbilt University Medical Center*

Christopher Fonnesbeck, *Vanderbilt University Medical Center*

Pingsheng Wu, *Vanderbilt University Medical Center*

Electronically-held medical databases containing information on the magnitude of hundreds of thousands to millions of records provide cost-effective resources to conduct observational cohort study. One class of medical administrative data that has found popularity among medical researchers is medical claims data. Motivated by a large scale observational cohort study with medical claims data, we develop a scalable Bayesian framework to accommodate time to first event of multivariate survival outcomes with ordinal severity. We model the multivariate survival outcomes using a flexible gamma frailty transformation model and provide a systematic and flexible way to determine the overall direction of the effect size using an additional data source correlating the type of survival outcomes with ordinal severity scores. A computationally efficient algorithm based on variational inference is used to scale the Bayesian inferential scheme to large datasets. Bayesian model selection procedures are further developed to determine the most proper and pragmatic transformation. Extensive numerical simulations are conducted to evaluate the validity of the method and variational algorithm. The proposed method is further applied to the Tennessee Asthma Bronchiolitis Study to investigate the maternal smoking effect on the childhood bronchiolitis.

✉ cindy.chen@vanderbilt.edu

RETROSPECTIVE STUDY DESIGNS FOR LONGITUDINAL DATA OBTAINED FROM A BIOBANK-LINKED ELECTRONIC MEDICAL RECORDS

Jonathan S. Schildcrout*, *Vanderbilt University Medical Center*

Retrospective, outcome dependent sampling (ODS) designs are efficient compared to standard designs because sampling is targeted towards those who are particularly informative for the estimation target. In the longitudinal data setting, one may exploit outcome vector, and possible covariate data, from an existing resource such as an electronic medical record to identify those whose expensive to ascertain and sample size limiting biomarker / exposure should be collected. Who is most informative is reasonably predictable and will depend upon the target of inference. In this talk, we will describe the class of designs, examine finite sampling operating characteristics, and apply the designs to an exemplar study on the impact of a single nucleotide polymorphism on cholesterol levels in patients taking statins at our institution.

✉ jonathan.schildcrout@vanderbilt.edu

101. ADAPTIVE DESIGN/ADAPTIVE RANDOMIZATION IN CLINICAL TRIALS

A NOVEL BAYESIAN PHASE I/II DOSE/SCHEDULE-FINDING DESIGN BASED ON BIOLOGICAL MECHANISM

Xiao Su*, *University of Texas MD Anderson Cancer Center*

Yisheng Li, *University of Texas MD Anderson Cancer Center*

We propose a Bayesian phase I/II dose/schedule-finding design to identify the biologically optimal dosing regimen, defined as the dose-schedule combination with the highest desirability in the risk-benefit tradeoff. The primary toxicity endpoint is time to dose-limiting toxicity (DLT) event and the primary efficacy endpoint is repeated measured solid tumor volume. We model the drug plasma concentration, toxicity endpoint and efficacy endpoint based on underlying biological mechanism. Firstly, the pharmacokinetics model is applied to establish the drug concentration trajectory over time. Conditional on the drug concentration trajectory, we model the solid tumor volume dynamics utilizing an ordinary differential equation. We also construct the hazard function for time to DLT by the Emax model condition on drug concentration. Using the accumulating data, we adaptively randomize patients to experimental doses based on the continuously updated model estimates. Simulation study, which shows that proposed design has good operating characteristics in terms of selecting the target dosing regimen and allocating patients to the optimal dose.

✉ xsu2@mdanderson.org

ADAPTIVE DESIGNS IN MULTI-READER MULTI-CASE CLINICAL TRIALS OF IMAGING DEVICES

Zhipeng Huang*, *U.S. Food and Drug Administration*

Weijie Chen, *U.S. Food and Drug Administration*

Frank Samuelson, *U.S. Food and Drug Administration*

Lucas Tcheuko, *U.S. Food and Drug Administration*

Evaluation of medical imaging devices often involves clinical studies where multiple readers read images of multiple cases - MRMC. In MRMC studies, both readers and cases contribute to the uncertainty of the estimated diagnostic performance, which is often measured by the area under the ROC curve (AUC). Due to limited prior information, the sizing of such a study is often unreliable and it is desired to adaptively re-size the study towards a target power after an interim analysis. Although adaptive design methods are available in clinical trials where only the patient sample is sized, such methodologies have not been established for MRMC studies. The challenge lies in the fact that there is a correlation structure in MRMC data and the sizing involves both readers and cases. We develop adaptive MRMC design methodologies to fill the gap. In particular, we resize the readers and cases to achieve a target power while adjusting the critical value for hypothesis testing to control the type I error in comparing AUCs of two modalities. Analytical results have been rigorously derived. Simulations show that the type I errors are controlled under a variety of simulation conditions.

✉ huangzp2016@gmail.com

A BAYESIAN PRECISION MEDICINE DESIGN FOR PHASE II/III CLINICAL TRIALS WITH MULTIPLE TREATMENTS

Liangcai Zhang*, *Rice University*

Suyu Liu, *University of Texas MD Anderson Cancer Center*

Ying Yuan, *University of Texas MD Anderson Cancer Center*

In cancer treatment practices, personalized medicine has been raised great awareness in response to cancer heterogeneity, or genetic diversity within a single cancer

population. In this article, we propose a bayesian personalized design to find customized treatments that cause fewer side effects but much more effectiveness to each patient according to his/her biomarker makeup/profile. Our model captures the biomarker-treatment and their interaction information that is predictive to the observed phase II/III outcomes. We first define the acceptable treatment sets based on the short term binary efficacy and longitudinal outcomes, and then identify the most effective treatment by monitoring patients' long-term efficacy endpoints. A two-stage treatment identification algorithm is proposed to find the personalized optimal treatment for patients with a specific biomarker pattern. Simulation studies show that the proposed design has higher probability of identifying the true treatment strategy than some existing designs.

✉ zhangliangcai2008@gmail.com

COMPARATIVE REVIEW OF TOXICITY PROBABILITY INTERVAL DESIGNS FOR PHASE I CLINICAL TRIALS

Heng Zhou*, *University of Texas MD Anderson Cancer Center*

Thomas A. Murray, *University of Minnesota*

Haitao Pan, *St. Jude Children's Research Hospital*

Ying Yuan, *University of Texas MD Anderson Cancer Center*

Recently, a number of new designs have been proposed that aim to combine the simplicity of algorithm-based designs with the superior performance of model-based designs, including the modified toxicity probability interval (mTPI), Bayesian optimal interval (BOIN) and Keyboard designs. In this article, we review these "model-assisted" interval designs, and compare their operating characteristics to the continual reassessment method (CRM). To provide more complete and reliable results, our comparison is based on 1000 dose-toxicity scenarios randomly generated using the pseudo-uniform algorithm recently proposed in the literature. The results showed that the CRM, BOIN and Keyboard designs provide comparable, excellent operating characteristics, and each outperforms the mTPI design. These designs are more likely to correctly select the MTD and less likely to overdose a large percentage of patients.

✉ hengzhou89@gmail.com

GROUP-SEQUENTIAL STRATEGIES IN CLINICAL TRIALS WITH BIVARIATE TIME-TO-EVENT OUTCOMES

Toshimitsu Hamasaki*, *National Cerebral and Cardiovascular Center*

Scott Evans, *Harvard School of Public Health*

Tomoyuki Sugimoto, *Kagoshima University*

Koko Asakura, *National Cerebral and Cardiovascular Center*

We discuss logrank test-based methods for efficacy or futility evaluation in group-sequential clinical trials that compare two interventions with respect to two time-to-event outcomes. Evaluation is conducted under two situations: (a) both events are non-composite, but one event is fatal, and (b) one event is composite, but other is fatal and non-composite. Based on group-sequential boundaries, we consider several decision-making frameworks for evaluating efficacy or futility. We consider two inferential goals, evaluating if a test intervention is superior to a control intervention on: (i) both outcomes (Co-primary endpoints: CPE), and (ii) at least one outcome (multiple primary endpoints: MPE). For the CPE goal, we incorporate the correlations among the outcomes into the calculations for non-binding futility boundaries and sample sizes as a function of other design parameters, including mean differences, the number of analyses, and efficacy boundaries. We investigate the operating characteristics of the decision-making frameworks in terms of power, the Type I error, sample sizes and event numbers.

✉ toshi.hamasaki@ncvc.go.jp

I02. FUNCTIONAL REGRESSION MODELS

BAYESIAN ANOVA MODELING FOR FUNCTIONAL DATA

Yu Yue*, *Baruch College, The City University of New York*

David Bolin, *Chalmers University of Technology, Sweden*

Havard Rue, *Norwegian University of Science and Technology, Norway*

Xiao-Feng Wang, *Cleveland Clinic Lerner Research Institute*

In this paper, we investigate the functional two-way ANOVA models from a novel Bayesian perspective. A class of highly flexible Gaussian Markov random fields (GMRF) are taken as priors on the functions in the model, which allows us to model various types of functional effects, such as (discrete or continuous) temporal effects and (point-level or areal) spatial effects. The resulting posterior distributions are obtained by an efficient computational tool based on integrated nested Laplace approximations (INLA) (Rue et al., 2009). We then employ the excursion method introduced by Bolin and Lindgren (2015) to build simultaneous credible intervals of functional effects and test their significance from a Bayesian point of view. A simulation study and multiple real data examples are presented to demonstrate the merits of our method.

✉ yu.yue@baruch.cuny.edu

MULTIVARIATE FUNCTIONAL RESPONSE REGRESSION

Hongxiao Zhu*, *Virginia Tech*

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

Fengrong Wei; *University of West Georgia*

Dennis D. Cox, *Rice University*

Many scientific studies measure different types of high-dimensional signals or images from the same subject, producing multivariate functional data. A joint analysis that integrates information across them may provide new insights into the underlying mechanism. Motivated by fluorescence spectroscopy data in a cervical pre-cancer study, a multivariate functional response regression model is proposed, which treats multivariate functional observations as responses and a common set of covariates as predictors. This novel modeling framework simultaneously accounts for correlations between functional variables and potential multi-level structures. The model is fitted by a two-stage linear transformation---a basis expansion to each functional variable followed by principal component analysis for the concatenated basis coefficients. This transformation effectively reduces the intra- and inter-function correlations and facilitates fast and convenient calculation. A fully Bayesian approach is adopted to sample the model parameters in the transformed space, and posterior inference is performed after inverse-transforming the regression coefficients back to the data domain.

✉ hongxiaozhu@vt.edu

TWO SAMPLE TESTS FOR LONGITUDINAL FUNCTIONAL DATA

Saebitna Oh*, *North Carolina State University*

Arnab Maity, *North Carolina State University*

Ana-Maria Staicu, *North Carolina State University*

We consider a model where longitudinal functional responses are observed at multiple time points per subject in two groups with additional scalar covariates. In this context, we develop a two sample testing procedure to test for equality of the time-varying covariate effect. Our test is based on a L2-norm based test statistic, and allows for the two groups to have different dependence structures. The size and power properties are examined in simulation studies. The methodologies?are illustrated with motivating real data example.

✉ soh3@ncsu.edu

GENERALIZED FUNCTIONAL LINEAR MODELS IN THE PRESENCE OF MISSING DATA WITH APPLICATION TO A RENAL STUDY

Will Zhu*, *Emory University*

Qi Long, *University of Pennsylvania*

Amita Manatunga, *Emory University*

Motivated by a renal study on detection of kidney obstruction, we consider an analysis of functional covariates and a binary outcome where up to two functional curves are observed for each subject. The second curve is sometimes missing and missingness is considered to be informative. We employ a Bayesian hierarchical model to jointly model the curves that are measured with error and the outcome, in which the association between noise-free curves and the outcome is of interest. We consider two approaches of selecting basis for modeling the curves and for parameterizing functional coefficients in the model for association. In the first approach, we use cubic B-spline basis functions and use deviance information criterion to select number of basis. As an alternative approach, we use functional principal component analysis (FPCA) to derive a more parsimonious model within the same framework to select the basis. We conduct simulation studies to assess the performance of the proposed methods. The simulation results demonstrate using basis from FPCA achieves better performance than using B-spline basis. The methods are applied to the motivating renal study.

✉ wzhu4@emory.edu

FUNCTIONAL VARIABLE SELECTION IN LOW-DIMENSIONAL INTERNAL MUSCULOSKELETAL BIOMECHANICAL (LIMB) MODEL

Md Nazmul Islam*, *North Carolina State University*

Jonathan Stallings, *North Carolina State University*

Ana-Maria Staicu, *North Carolina State University*

He Huang, *North Carolina State University*

The broad objective of this project is to develop a novel low-dimensional, electromyography (EMG) controlled, multifunctional robotic prosthetic limb for transradial amputee. To accomplish this task, identifying the influential EMG signals accounting for the neuromuscular and biomechanical aspects plays a significant role. We consider functional linear regression to assess the systematic association between signals measuring muscle contractions and hand-wrist movements across different postures of able-bodied subjects. We propose a two-step parsimonious modeling framework; first select important variables using an extension of adaptive group-lasso regularization technique, next refit the model on the selected subset to reduce shrinkage bias. Our method of variable selection facilitates to identify the clinically important EMG signals with negligible false positive rates. Our proposed methodology is also applicable in a general setting where the functional coefficients are varying over covariates. Extensive simulation study on high-dimensional data shows excellent numerical performance in terms of variable selection with superior predictive performance.

✉ mnislam@ncsu.edu

I03. ANALYSIS OF ROC CURVES

A COVARIATE-ADJUSTED CLASSIFICATION MODEL FOR MULTIPLE BIOMARKERS IN DISEASE SCREENING AND DIAGNOSIS

Suizhi Yu*, *Kansas State University*

Wei-Wen Hsu, *Kansas State University*

The classification methods based on linear biomarker combinations have been widely used to improve the accuracy in disease diagnosis. However, it is seldom to include covariates such as gender and age at diagnosis into these classification procedures. It is known that biomarkers or patient outcomes are often associated with covariates in practice, therefore the inclusion of covariates in the model may further increase the power of prediction and improve the classification accuracy. In this paper, a covariate-adjusted classification model for multiple biomarkers is pro-

posed. Technically, it is a two-stage model with a parametric or non-parametric approach to combine biomarkers first, and then incorporating covariates with the use of maximum rank correlation estimators. Specifically, these parameter coefficients associated with covariates can be estimated by maximizing the area under the ROC curve. The asymptotic properties of these estimators in the model are provided and an intensive simulation study is conducted to evaluate the performance in finite sample sizes. The data of colorectal cancer are used to illustrate the proposed methodology.

✉ smallcowzhi@gmail.com

A MODEL-FREE FRAMEWORK TO DETERMINING THE COVARIATE-ADJUSTED YOUDEN INDEX AND ITS ASSOCIATED CUT-POINT

Jiwei Zhao*, *State University of New York at Buffalo*

In medical research, the accuracy of diagnostic tests is commonly assessed using the receiver operating characteristic (ROC) curve. To summarize an ROC curve and determine its optimal cut-point, the Youden index is popularly used. In literature, the estimation of the Youden index has been widely studied via various statistical modeling strategies on the conditional density. This paper proposes a novel model-free framework, which directly estimates the covariate-adjusted cut-point without estimating the conditional density. The proposed method formulates the estimation problem in a large margin classification framework, and the variable selection techniques are employed to facilitate determining the covariate-adjusted Youden index and its associated cut-point. The advantage of the proposed method is demonstrated in a variety of simulated experiments as well as a real application. The connection of our method to the concept of minimal clinically important difference (MCID) is also discussed.

✉ jiwei2012zhao@gmail.com

NEW APPLICATIONS OF KULLBACK-LEIBLER INFORMATION AS A MEASURE OF DIAGNOSTIC ACCURACY AND ITS RELATION TO COMMON ROC INDICES

Jingjing Yin*, *Georgia Southern University*

Hani Samawi, *Georgia Southern University*

Xinyan Zhang, *Georgia Southern University*

Lili Yu, *Georgia Southern University*

Haresh Rochani, *Georgia Southern University*

Robert Vogel, *Georgia Southern University*

The Receiver Operating Characteristic curve (ROC) is used for evaluating and comparing diagnostic tests. The Kullback-Leibler distance measure (D), which captures the disparity between two distributions, can be considered as one of the indices for determining the diagnostic performance of markers. This study presents some new applications of Kullback-Leibler distance in medical diagnostics, including overall measure of before-test rule-in and rule-out potential as well as an optimization criteria for cut point selection. Furthermore, the paper links Kullback-Leibler distance with some common ROC measures and demonstrates analytically and numerically the relations in situations of one cut point as well as multiple cut points. Moreover, a graphical application and interpretation of Kullback-Leibler distance, which is referred as the information graph is discussed. Numerical examples as well as a real data analysis are provided to illustrate the proposed new applications.

✉ jyin@georgiasouthern.edu

MEASURING ACCURACY OF BIOMARKERS UNDER TREE OR UMBRELLA ORDERING

Yingdong Feng*, *State University of New York at Buffalo*

Lili Tian, *State University of New York at Buffalo*

Measuring accuracy of biomarkers is essential in many applied fields. For diseases with multi-classes, an important category is tree/umbrella ordering in which the marker measurement for one particular class is lower/higher than those for the rest classes. For example, non-small cell lung cancer (NSCLC) is a collective term for several subtypes among which no clearly defined order exist. Therefore, the problem of identifying NSCLC from either small cell lung cancer (SCLC) or healthy group falls into the framework of tree/umbrella ordering. The recently proposed TROC curve and TAUC are extensions of ROC curve and AUC for tree/umbrella ordering. In this talk, we explore a new diagnostic measure which has been neglected for tree/umbrella ordering, namely integrated false negative rate (ITFNR), and propose the idea of using both TAUC and ITFNR to evaluate the diagnostic accuracy of a biomarker under tree/umbrella ordering. Parametric and non-parametric approaches for constructing joint confidence region are proposed. Simulation studies are carried out to assess and compare the performance of these methods, and a published microarray data set is analyzed.

✉ yfeng6@buffalo.edu

STATISTICAL INFERENCE OF TWO CLASSIFIERS BY AREA UNDER THE ROC CURVE WITH EMPIRICAL LIKELIHOOD

Xue Ding*, *University of Kentucky*

Mai Zhou, *University of Kentucky*

The Receiver Operating Characteristic (ROC) curve has been extensively used in the assessment of medical diagnostic tests. As a quantitative summary index, the area under ROC curve (AUC) measures the overall accuracy of classifying the diseased subjects from non-diseased subjects. Using two-sample empirical likelihood methodology by Owen (2001), we investigate the difference between the areas under two ROC curves with paired data. Unlike previously proposed methods, the empirical likelihood ratio in our study is asymptotically chi square distributed without any adjustment. Moreover, compared to the existing nonparametric approaches, our test procedure avoids having to estimate the variance, resulting in a more accurate test statistic. In addition, the corresponding confidence interval is invariant to transformation. We illustrate our approach in a real data example and evaluate its performance in the simulation studies.

✉ xdi226@g.uky.edu

ASSESSING PREDICTIVE VALUE OF RISK FACTORS IN LARGE-SCALE PROSPECTIVE OBSERVATIONAL STUDIES

Xiang Liu*, *University of South Florida*

Kendra Vehik, *University of South Florida*

Jeffrey Krischer, *University of South Florida*

In large-scale prospective observational studies, new risk factors are discovered over time during the long-term follow-up. As knowledge accumulates, the emergence of a new risk factor leads to an important question about how much prediction accuracy the new risk factor achieves and how much improvement in prediction accuracy it is achieved by adding the new marker into the existing set of risk factors. A systematic analytic framework has been proposed (i) to construct risk prediction models using variable selection methods for Cox proportional hazard models; (ii) to access the discrimination ability of each prediction model using time-dependent ROC curve analysis; and (iii) to visualize the incremental value of each risk factor for prediction of disease. Variable selection procedures including stepwise selection and regularization are considered to address the statistical challenge of dealing with a large number of risk factors (high-dimensionality) and also to address the practical demand for presenting the incremental value of each risk factor. Real data will be used to illustrate and compare the performance of different variable selection procedures.

✉ xiang.liu@epi.usf.edu

I04. JOINT MODELS FOR SURVIVAL AND LONGITUDINAL DATA

EFFECT OF ADHERENCE ON AIDS RELATED OUTCOMES IN HIV PATIENTS – A LIKELIHOOD BASED APPROACH TO CORRECT FOR MISCLASSIFICATION IN A COX PROPORTIONAL HAZARDS MODEL

Varadan Sevilimedu*, *Yale School of Public Health*

Shuangge Ma, *Yale School of Public Health*

Tassos Kyriakides, *Department of Veteran Affairs*

Our study is based on a clinical trial that evaluates the effectiveness of a novel anti-retroviral therapeutic (ART) regimen in treating patients affected by the Human Immunodeficiency virus (HIV). Specifically, we study the effect (on parameter estimates) of the discrepancy between self-reported adherence to the novel ART regimen and true adherence obtained using validation data. This discrepancy between self-reported adherence and true adherence can lead to misclassification error and hence produce attenuated parameter estimates of the misclassified covariate a in Cox proportional hazards model. We propose likelihood based methods to correct for this attenuation caused by misclassification in time varying covariates in a Cox regression model. The motivating data for this study comes from information collected on 500 patients suffering from HIV, in whom regular ART therapy was not effective. We expect to obtain significantly more accurate estimates of the effect of the novel ART regimen on the hazards of AIDS related events, when the estimates are corrected for misclassification error using the proposed method.

✉ varadan.sevilimedu@yale.edu

JOINT ANALYSIS OF MULTIPLE LONGITUDINAL PROCESSES WITH SKEWNESS AND EXCESS OF ZEROES AND A DISCRETE SURVIVAL TIME: AN APPLICATION TO FECUNDITY DATA

Sedigheh Mirzaei Salehabadi*, *Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health*

Somak Chatterjee, *The George Washington University*

Subrata Kundu, *The George Washington University*

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

We consider the joint modeling, analysis and prediction of multiple longitudinal processes and discrete time to event. Our motivation comes from interest in understanding the patterns of multiple phases of menstrual cycle length, and their association with time to pregnancy. The two phases of menstrual cycle length, separated by ovulation have tremendous inter- and intra woman variability, which are also influenced by a woman's risk factors. It is of considerable scientific interest to study the patterns of these processes in reproductive aged women, who are attempting to conceive. Motivated by these interests, we study the joint modeling of follicular phase, luteal phase and their association with time-to-pregnancy. We propose a mixture distribution with normal and Gumbel distribution for modeling the follicular phase and propose a log-normal distribution with excess of zeroes to model the luteal phase to account for cycles where a woman does not ovulate and hence has no luteal phase. Additionally, a shared parameter model is used to capture the dependence between the mean lengths of follicular and luteal phase with time-to-pregnancy.

✉ mirzaeisalehas@mail.nih.gov

APPLYING SURVIVAL ANALYSIS AND COUNT MODELS TO TWITTER DATA

Congjian Liu*, *Georgia Southern University*

Jingjing Yin, *Georgia Southern University*

Chun Hai (Isaac) Fung, *Georgia Southern University*

Lindsay Mullican, *Georgia Southern University*

Twitter has a variety of information on it, health topic is one of the popular categories. We used a collection of almost 40,000 tweets extracted from Twitter with #blood pressure from January, 2014 to April, 2015 to investigate the potentially associated factors for popularity (measured by the number of retweet) as well as the survival of tweets (measured by the time frame from the first post to its last retweet). We have found the appearance of a few hashtags significantly decreased the survival of tweets. Furthermore, these hashtags increase(but some decrease) the odds of being retweeted. And other factors significantly associated with the odds include actor's friends count, actor's follower's count, actor's listed count and so on. We explored our results using R, the results do not highlight the potential of hashtag in the application of twitter.

✉ cl03124@georgiasouthern.edu

ON THE LANDMARK SURVIVAL MODEL FOR DYNAMIC PREDICTION OF EVENT OCCURRENCE USING LONGITUDINAL DATA

Yayuan Zhu*, *University of Texas MD Anderson Cancer Center*

Liang Li, *University of Texas MD Anderson Cancer Center*

Xuelin Huang, *University of Texas MD Anderson Cancer Center*

In longitudinal cohort studies, participants are often monitored through periodic clinical visits until the occurrence of a terminal clinical event. A question of interest to both scientific research and clinical practice is to predict the risk of the terminal event at each visit, using the longitudinal prognostic information collected up to the visit. This problem is called the dynamic prediction: a real-time, personalized prediction of the risk of a future adverse clinical event with longitudinal prognostic information. We first review the landmark modeling approaches to dynamic prediction problems, with a focus on the landmark Cox model. A challenge in the methodological research of the landmark Cox model is to generate a dataset which satisfies infinitely many Cox models at any landmark time. In current literature, the statistical properties of landmark prediction models are often studied using data simulated from joint models. We propose an algorithm to generate data from the assumed landmark model directly so as to avoid studying the model under misspecification. It will facilitate future theoretical and numerical research on landmark dynamic prediction models.

✉ yzhu8@mdanderson.org

I05. BIOMARKERS

ROBUST ΔΔCT ESTIMATE

Arun Kumar*, *Livanova USA Inc.*

Daniel Lorand, *Novartis Pharma AG*

△△ct method estimates fold change (in log base 2 scale) in a gene expression for data coming from RT-PCR assay. The △△ct estimate aggregates replicates using mean and standard deviation (sd) and hence is not robust to outliers which are in practice often removed before the non-outlying replicates are aggregated. Subjective removal of outliers leads to variation in results from user to user. Alternative is to use robust statistics such as median and median absolute deviation (MAD) to aggregate the replicates but is not done in practice perhaps because the distribution of a robust △△ct estimate based on median and MAD is not straightforward and hence inference for such robust △△ct estimate is challenging for practitioners. In this presentation, we use a statistical modeling framework to deduce an approximate distribution for a robust △△ct estimate. Simulation results are presented to show that the robust △△ct estimate compared to △△ct estimate used in practice leads to significantly reduced confidence interval length when data has outliers.

✉ m25arun@gmail.com

ASSOCIATION OF BIOMARKERS WITH PROGRESSIVE DISEASE STATES

Julia E. Crook (Kelsall)*, *Mayo Clinic*

In Alzheimer's disease (AD) research, it is typically assumed that a good biomarker of AD development will have increasing or decreasing levels as an individual moves along the spectrum of disease from being cognitively normal to cognitively impaired and then on to AD. For a simple comparison of ADs and controls, a straightforward linear regression adjusting for relevant covariates, such as age, sex, APOE genotype, can usually be used for analysis. When there are intermediate states between cognitively normal and AD, the most appropriate analysis is less clear. A constrained likelihood version of linear regression is considered that constrains the mean biomarker levels to be monotonically increasing or decreasing with worsening disease state. The usual F-statistic can be constructed and it turns out that, under the usual assumptions of linear regression, it is distributed as a weighted mixture of F distributions with weights dependent on the number of patients in each of the ordered categories. Naturally, this approach provides increased power to detect truly monotonic associations as compared to inclusion of disease state as an unordered categorical variable.

✉ crook.julia@mayo.edu

A LATENT CLASS APPROACH FOR JOINT MODELING OF A TIME-TO-EVENT OUTCOME AND MULTIPLE LONGITUDINAL BIOMARKERS SUBJECT TO LIMITS OF DETECTION

Menghan Li*, *The Pennsylvania State University*

Lan Kong, *The Pennsylvania State University*

Multiple biomarkers on different biological pathways are often measured over time to investigate the complex mechanism of disease development and progression. Identification of informative subpopulation patterns of longitudinal biomarkers and clinical endpoint may greatly facilitate risk stratification and decision-making of treatment strategies. We develop a joint latent class model for multiple biomarkers and a time-to-event outcome while accounting for the censored biomarker measurements due to limits of detection. Latent class modeling captures the interrelationship between biomarker trajectories and clinical endpoint via latent classes, which reveal the subpopulation profiles of biomarkers and clinical outcome. The estimation of joint latent class models can be complicated by censored biomarker measurements, we provide a Monte Carlo EM (MCEM) algorithm for maximum likelihood estimation. We demonstrate the satisfactory performance of our MCEM algorithm using simulation studies and apply the proposed method to a motivating study to discover longitudinal profiles of cytokine responses to pneumonia and their associated mortality risk.

✉ mul283@psu.edu

NONPARAMETRIC CONDITIONAL DENSITY ESTIMATION FOR BIOMARKERS BASED ON POOLED ASSESSMENTS

Xichen Mou*, *University of South Carolina*

Dewei Wang, *University of South Carolina*

Joshua M. Tebbs, *University of South Carolina*

The process of assaying pooled specimens is becoming increasingly common in evaluating the diagnostic efficacy of biomarkers. It has been shown that a biomarker's diagnostic efficacy often depends on individual covariate information. In order to better understand this dependence, we present a distribution-free method to estimate the conditional density of a biomarker level given a continuous covariate where the biomarker levels are measured subject to pooling and potential measurement errors. Two different types of pooling strategies are considered: i.e., random pooling, which pools individual specimens randomly, and homogeneous pooling, where specimens of individuals that have similar covariates are pooled together. Under each strategy, we derive the asymptotic properties of the new density estimator. Simulation studies demonstrate that our approaches can accurately estimate the conditional density in a variety of settings. We further illustrate the new methodology by applying it to a diabetes dataset.

✉ xmou@email.sc.edu

A BAYESIAN SCREENING APPROACH FOR HEPATOCELLULAR CARCINOMA USING MULTIPLE LONGITUDINAL BIOMARKERS

Nabihah Tayob*, *University of Texas MD Anderson Cancer Center*

Francesco Tayob, *University of Florence*

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, therefore early detection is critical to improving the survival of patients with HCC. Serum τ-Fetoprotein (AFP) is widely used, but it has limited sensitivity and is not elevated in all HCC cases so, we incorporate a second blood-based biomarker, des-τcarboxy-prothrombin (DCP), that has shown potential. 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 for HCC. We assume the trajectories of AFP and DCP follow a joint hierarchical mixture model with random changepoints that allows for distinct changepoint times and subsequent trajectories of each biomarker. Changepoint indicators are jointly modeled with a Markov Random Field distribution to help detect borderline changepoints. Markov chain Monte Carlo methods are used to calculate posterior distributions, which are used in risk calculations among future patients and determine whether a patient has a positive screen. The screening algorithm was compared to alternatives in the HALT-C Trial using cross-validation.

✉ ntayob@gmail.com

I06. BAYESIAN METHODS FOR GENETICS AND GENOMICS

NEGATIVE BINOMIAL BAYESIAN GENERALIZED LINEAR MODEL FOR ANALYZING HIGH-DIMENSIONAL OTU COUNT DATA

Amanda H. Pendegraft*, *University of Alabama at Birmingham*

Nengjun Yi, *University of Alabama at Birmingham*

Boyi Guo, *University of Alabama at Birmingham*

It is well-known that modern next-generation sequencing technology has allowed researchers to collect large volumes of metagenomic sequencing data. A common example, 16S rRNA gene amplicon sequencing, identifies operational taxonomic units (OTUs) so to enable the study of bacterial community composition in association with environmental, clinical, and demographic host characteristics. However, current analytical tools restrict study designs to investigations of a limited number of the respective host characteristics which may yield a loss of relevant information. Efficient statistical techniques are therefore needed to compensate for high-dimensional combinations involving tens or hundreds of host characteristics further expanding the analysis and interpretation of microbiome count data. In this presentation, we propose a hierarchical Negative Binomial Bayesian generalized linear model capable of simultaneously adjusting for complex study designs and provide an illustration our method using demographic and dietary responses of 824 "healthy" American Gut Project participants.

✉ alhall91@uab.edu

KNOWLEDGE-GUIDED BAYESIAN VARIABLE SELECTION IN SUPPORT VECTOR MACHINES FOR STRUCTURED HIGH-DIMENSIONAL DATA

Wenli Sun*, *University of Pennsylvania*

Changgee Chang, *University of Pennsylvania*

Qi Long, *University of Pennsylvania*

Support vector machines (SVM) is a popular classification method for analysis of high dimensional data such as genomics data. Recently, new SVM methods have been developed to achieve feature selection through either frequentist regularization or Bayesian shrinkage. The Bayesian SVM framework proposed by Law and Kwok (2001) provides a probabilistic interpretation for SVM and allows direct uncertainty quantification. Building on this framework, we propose a new Bayesian SVM method that enables feature selection guided by structural information among predictors, e.g, biological pathways among genes. Our method uses a spike and slab prior for feature selection combined with a markov random filed prior for incorporating structural information. Markov chain Monte Calro algorithm is developed for a full Bayesian inference. The performance of our method is evaluated in comparison with existing SVM methods in terms of prediction and feature selection in extensive simulations. Our SVM method is also illustrated in analysis of genomic data from a cancer study, demonstrating its advantage in generating biologically meaningful results and identifying potentially important features.

✉ sunwenli1234@gmail.com

THE SPIKE-AND-SLAB LASSO COX MODEL FOR SURVIVAL PREDICTION AND ASSOCIATED GENES DETECTION

Xinyan Zhang*, *Georgia Southern University*

Zaixiang Tang, *Soochow University*

Yueping Shen, *Soochow University*

Nengjun Yi, *University of Alabama at Birmingham*

Motivation: Large-scale molecular profiling data have offered extraordinary opportunities to improve survival prediction of cancers and other diseases and to detect disease associated genes. However, there are considerable challenges in analyzing large-scale molecular data. Results: We propose new Bayesian hierarchical Cox proportional hazards models, called the spike-and-slab lasso Cox, for predicting survival outcomes and detecting associated genes. We also develop an efficient algorithm to fit the proposed models by incorporating Expectation-Maximization steps into the extremely fast cyclic coordinate descent algorithm. The performance of the proposed method is assessed via extensive simulations and compared with the lasso Cox regression. We demonstrate the proposed procedure on two cancer datasets with censored survival outcomes and thousands of molecular features. Our analyses suggest that the proposed procedure can generate powerful prognostic models for predicting cancer survival and can detect associated genes. Availability and implementation: The methods have been implemented in a freely available R package BhGLM (http://www.ssg.uab.edu/bhglm/).

✉ xzhang@georgiasouthern.edu

A BAYESIAN FRAMEWORK FOR REWIRING THE TOPOLOGICAL NETWORK OF INTRATUMORAL CELLS

Lin Qiu*, *The Pennsylvania State University*

Vernon M. Chinchilli, *The Pennsylvania State University*

Rongling Wu, *The Pennsylvania State University*

WITHDRAWN

Intratumoral heterogeneity (ITH) has been regarded as a key cause of the failure and resistance of cancer therapy, but how it behaves and functions remains unclear. Advances in single-cell analysis have facilitated the collection of a massive amount of data about genetic and molecular states of individual cancer cells, providing a fuel to dissect the mechanistic organization of ITH at the molecular, metabolic and positional level. Taking advantage of these data, we develop a Bayesian model to rewire up a topological network of cell-cell interdependences and interactions that operate within a tumor mass. The model is grounded on the premise of game

theory that each interactive cell (player) strives to maximize its fitness by pursuing a "rational self-interest" strategy in a way that senses and alerts other cells to responds properly. By integrating this idea with genome-wide association studies for intratumoral cells, the model is equipped with a capacity to visualize, annotate and quantify how somatic mutations mediate ITH and the network of intratumoral interactions.

✉ lquva@utexas.edu

BAYESIAN INDICATOR VARIABLE SELECTION MODEL TO INCORPORATE MULTI-LAYER OVERLAPPING GROUP STRUCTURE IN MULTI-OMICS APPLICATIONS

Li Zhu* •, *University of Pittsburgh*

Zhiguang Huo, *University of Florida*

Tianzhou Ma, *University of Pittsburgh*

George Tseng, *University of Pittsburgh*

Variable selection is a pervasive question in modern high-dimensional data analysis. Incorporation of group structure knowledge to improve variable selection has been widely studied. Here, we consider prior knowledge of a multi-layer overlapping group structure to improve variable selection in regression setting. In genomic applications, for instance, a biological pathway contains tens to hundreds of genes and a gene can contain multiple experimentally measured features. In addition to the multi-layer structure, the groups may be overlapped. Incorporating such hierarchical multi-layer overlapping groups in traditional penalized regression setting produces difficulty in optimization. In this paper, we propose a Bayesian indicator model that can elegantly serve the purpose. We discuss the soft-thresholding property of the posterior median estimator and prove its selection consistency and asymptotic normality under orthogonal design. We apply the model to simulations and two breast cancer examples to demonstrate its superiority over other existing methods. The result not only enhances prediction accuracy but also improves variable selection and model interpretation.

✉ liz86@pitt.edu

107. UNCOVERING HETEROGENEITY IN LONGITUDINAL DATA: CLUSTERING AND MIXTURE MODELING

TREE-BASED CLUSTERING OF LONGITUDINAL CHILDHOOD GROWTH

Brianna Heggeseth*, *Williams College*

There is variation in adiposity growth among children in the United States. We seek to characterize the heterogeneity in growth patterns of childhood body mass index and explore possible associations with many early-life factors. There is a growing literature to suggest that early-life exposure to a mixture of chemicals may increase the risk of unhealthy obesity development by disrupting hormonal processes that mediate growth, potentially explaining some variation in growth. To explore relationships with multipollutant exposures, we propose utilizing tree-based methods for finding children with similar growth patterns and similar exposure levels. We discuss extensions of the CART algorithm to accomodate longitudinal data and modifications that define similarity in terms of growth pattern. We illustrate how this approach allows for the possible discovery of complex interactions between chemical exposures as well as non-linear associations.

✉ bch2@williams.edu

CLUSTERING DISCRETE STATE TRAJECTORIES OF VARYING LENGTHS: HEALTH CARE UTILIZATION PATTERNS

Laura A. Hatfield*, *Harvard Medical School*

Megan S. Schuler, *RAND Corporation*

Nina R. Joyce, *Brown University*

Elizabeth B. Lamont, *Harvard Medical School*

Haiden A. Huskamp, *Harvard Medical School*

Understanding how health outcomes and care evolve over time is important in clinical and policy research. In our motivating example, we wish to understand heterogeneity in care for Medicare beneficiaries after diagnosis with a rapidly

fatal form of cancer. We want to group individuals with similar patterns of health care utilization. The individual health care utilization trajectories are sequences of varying length (depending on how long a person survives) that comprise a sequence of episodes of time spent in five mutually exclusive health care settings. Between diagnosis and death, individuals transition among care settings (hospital, post-acute nursing facility, hospice), spending different lengths of time in each setting. In this talk, I will present methods for examining heterogeneity in state transition sequences. We combine feature extraction and latent class analysis. We begin by characterizing features of health trajectory data, such as those that arise from administrative data, surveys, and ecological momentary assessments. We then discuss how latent class analysis can be used to cluster individuals with heterogeneous trajectories defined by multiple factors.

✉ hatfield@hcp.med.harvard.edu

CLUSTERING LONGITUDINAL DATA

Paul D. McNicholas*, *McMaster University*

Different approaches for clustering longitudinal data are presented. While each one is based on a mixture model, they differ in several respects. Some methods are designed to cluster a single variable over time, and are based on multivariate finite mixture models. These approaches are useful, for example, in gene expression time course studies. Methods based on mixtures of matrix variate distributions are also discussed. Such approaches facilitate cluster analyses that consider multiple random variables over time, and are useful for data from a wide range of studies. The option to allow for skewness or heavy tails in the analyses, as well as the possible use of latent variables, are also discussed.

✉ paulmc@mcmaster.ca

SEMI-PARAMETRIC MIXTURE MODELING OF LONGITUDINAL DATA: IMPLICATIONS FOR IDENTIFYING HETEROGENEOUS TREATMENT EFFECTS

Amelia M. Haviland*, *Carnegie Mellon University*

Hilary Wolfendale, *Carnegie Mellon University*

A central theme of research on disease progression is whether an intervention or treatment has different impacts depending on the trajectory a person was following prior to the intervention. In this work, we focus on observational study settings with individual longitudinal data in additional to baseline covariates. Semi-parametric finite mixture modeling is used to identify trajectories of the longitudinal data present in the data prior to treatment being administered. The trajectories serve three purposes, they describe heterogeneous progression over time, identify classes of subjects for whom no good matches are available, and suggest meaningful classes of subjects to test for heterogeneous treatment effects. Propensity score matching is applied to treated and untreated subjects, within trajectory, on baseline covariates. Balancing on prior values of the measure that will be used as the treatment outcome gives particular protection from unobserved confounders. The methods are applied to a population of chronic kidney disease patients for whom distinct trajectories of disease progression have recently been identified.

✉ haviland@cmu.edu

I08. STATISTICAL METHODS FOR CANCER -OMIC DATA

A MULTI-VIEW SPECTRAL CLUSTERING METHOD TO ANALYZE DIVERSE 'OMICS DATA SETS

Hongyu Zhao*, *Yale University*

Seyoung Park, *Yale University*

Hao Xu, *Sichuan University*

Advances in high-throughput genomic technologies and large-scale studies, such as The Cancer Genome Atlas (TCGA) project, have generated rich information from a wide range of -omics data sets for many cancer types. Clustering patients incorporating multiple -omics data types has the potential to better cluster patients, understand genetic heterogeneity, and predict patient's prognosis. In this presentation, we propose a novel multi-view spectral clustering method

which learns the weight of each data type and a similarity measure between patients based on multiple Gaussian kernels via a non-convex optimization framework. We solve the proposed non-convex problem iteratively using the ADMM algorithm and show the convergence of the algorithm. We evaluate the performance of the proposed clustering method on various simulated data to demonstrate its effectiveness and robustness. When our method is applied to 22 different cancer datasets, the inferred clusters are better correlated with patients' survival compared to other clustering methods. This is joint work with Seyoung Park and Hao Xu.

✉ hongyu.zhao@yale.edu

PRECISE - PERSONALIZED CANCER-SPECIFIC INTEGRATED NETWORK ESTIMATION

Kim-Anh Do*, *University of Texas MD Anderson Cancer Center*

Min Jin Ha, *University of Texas MD Anderson Cancer Center*

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

The functional cancer genomic and proteomic data provide rich sources of information to identify variations in signaling pathways and activities within and across tumor lineages. However, current analytic methods lack the ability to exploit the diverse and layered architecture of biological networks to provide coherent metrics for inferring pathway-based activities that can be used for both global cancer-specific and local patient-specific stratification and outcome prediction. We propose personalized cancer-specific integrated network estimation (PRECISE), a general framework for integrating existing interaction databases, data-driven de novo causal structures, and upstream molecular profiling data to estimate cancer-specific integrated networks, infer patient-specific networks and elicit interpretable pathway-level signatures. We develop a Bayesian regression model for protein-protein interactions that integrates with known pathway annotations and protein-protein interactions.

✉ kimdo@mdanderson.org

USING eQTLs TO DISCOVER NOVEL GENETIC LOCI FOR COMPLEX DISEASES

Li Hsu*, *Fred Hutchinson Cancer Research Center*

Wei Sun, *Fred Hutchinson Cancer Research Center*

Vicky Wu, *Fred Hutchinson Cancer Research Center*

The recent rapid development in collecting multiple types of genetic and genomics data on hundreds of samples has allowed for researchers to identify quantitative trait loci (QTLs) associated with gene expression or other types genomics features. The knowledge of these eQTLs can then be used for targeted, gene-based association analysis in the large-scale genetic studies of complex diseases. In this talk, I will describe some recent works on improving identification of eQTLs by accounting for gene-expression networks using penalized regression approach. I will also propose a statistical framework for incorporating the functional information from eQTLs in the down-stream genetic association analysis, and introduce score test statistics for jointly testing the association of eQTLs with disease risk. Simulation and real data analyses will be shown to demonstrate the performance of the proposed method.

✉ lih@fredhutch.org

I09. GEOMETRIC APPROACHES TO FUNCTIONAL DATA ANALYSIS FOR BIOMEDICAL APPLICATIONS

STATISTICAL SUMMARIZATION, PRINCIPAL MODES AND SHAPE MODELING FOR SIMPLIFIED NEURONAL TREES

Adam G. Duncan*, *Florida State University*

Eric Klassen, *Florida State University*

Anuj Srivastava, *Florida State University*

Neuron morphology plays a central role in characterizing cognitive health and functionality of brain structures. The problem of quantifying differences in neuron shapes and

capturing statistical variability of shapes is difficult because neurons differ both in geometry and in topology. We present a mathematical representation of neuronal trees, restricting to trees with a simplified class of topologies. We impose a metric space on the set of such trees based on elastic distances between curve shapes of individual branches, in order to compare neuronal shapes, and to obtain optimal deformations (geodesics) across arbitrary trees. A key part of this metric is to define certain equivalence relations that allow trees with different geometries and topologies to be compared efficiently by finding an optimal registration between corresponding parts of different trees. This registration method is generalized to simultaneously compare many trees, and find sample Fréchet means and modes of variability. The framework is illustrated on simulated trees and real datasets of neuron reconstructions which had previously been extracted from confocal microscope images.

✉ a.duncan@stat.fsu.edu

MODELING MULTI-WAY FUNCTIONAL DATA UNDER WEAK SEPARABILITY, WITH APPLICATION TO BRAIN FUNCTIONAL CONNECTIVITY

Kehui Chen*, *University of Pittsburgh*

Brian Lynch, *University of Pittsburgh*

Multi-way functional data refers to functional data with double or multiple, such as brain-imaging data with spatial and temporal indices. In practice, the number of spatial grids and the number of time grids both could be very large. To achieve efficient dimension reduction, one usually adopts the strong separability assumption that the covariance can be factorized as a product of a spatial covariance and a temporal covariance. This assumption is quite restrictive and is often violated in real applications. In the talk, we will introduce a new concept of weak separability, where the covariance can be approximated by a weighted sum of spatial-temporal separable components, including strong separability as a special case. Using this notion, we will discuss some analysis tools for brain functional connectivity based on MEG data.

✉ khchen@pitt.edu

SAMPLING WARPING FUNCTIONS FOR CURVE REGISTRATION

Karthik Bharath*, *University of Nottingham*

Sebastian Kurtek, *The Ohio State University*

Registration of functional or curve data in the presence of important landmark information is a frequently encountered task in several biomedical applications. We propose a sampling scheme for warp maps used in alignment of open and closed curves, possibly with landmark constraints. The scheme provides a point process-based constructive definition of a probability measure on the set of warp maps of [0, 1] and the unit circle. The measure is used as a prior on warp maps in a Bayesian model for alignment, and as a proposal distribution in a stochastic algorithm to solve a variational formulation of curve alignment.

✉ karthik.bharath@nottingham.ac.uk

RADIOLOGIC IMAGE-BASED STATISTICAL SHAPE ANALYSIS OF BRAIN TUMORS

Sebastian Kurtek*, *The Ohio State University*

Karthik Bharath, *University of Nottingham*

Arvind U.K. Rao, *University of Texas MD Anderson Cancer Center*

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

We propose a curve-based Riemannian-geometric approach for general shape-based statistical analyses of tumors obtained from radiologic images. A key component of the framework is a suitable metric that (1) enables comparisons of tumor shapes, (2) provides tools for computing descriptive statistics and implementing principal component analysis on the space of tumor shapes, and (3) allows for a rich class of continuous deformations of a tumor shape. The utility of the framework is illustrated through specific statistical tasks on a dataset of radiologic images of patients diagnosed with glioblastoma multiforme, a malignant

brain tumor with poor prognosis. In particular, our analysis discovers two patient clusters with very different survival, subtype and genomic characteristics. Furthermore, it is demonstrated that adding tumor shape information into survival models containing clinical and genomic variables results in a significant increase in predictive power.

✉ kurtek.1@stat.osu.edu

IIO. RANDOMIZATION INFERENCE: A BACK TO THE FUTURE PERSPECTIVE

RANDOMIZATION INFERENCE WITH GENERAL INTERFERENCE AND CENSORING

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

Wen Wei Loh, *Ghent University*

Interference occurs between individuals when the treatment (or exposure) of one individual affects the outcome of another individual. Previous work on causal inference methods in the presence of interference has focused on the setting where a priori it is assumed there is "partial interference" in the sense that individuals can be partitioned into groups wherein there is no interference between individuals in different groups. In this talk we consider randomization-based inferential methods proposed by Bowers et al. (2012, 2016) that allow for more general interference structures in the context of randomized experiments. Extensions of the Bowers et al. approach are considered, including allowing for right censored outcomes. The methods are utilized to assess whether interference is present in data from a cholera vaccine trial of n=73,000 women and children in Matlab, Bangladesh.

✉ mhudgens@bios.unc.edu

BEYOND THE SHARP NULL: RANDOMIZATION INFERENCE, BOUNDED NULL HYPOTHESES, AND CONFIDENCE INTERVALS FOR MAXIMUM EFFECTS

Luke W. Miratrix*, *Harvard University*

Devin Caughey, *Massachusetts Institute of Technology*

Allan Dafoe, *Yale University*

Fisherian randomization inference is often dismissed as testing an uninteresting and implausible hypothesis: the sharp null of no effects whatsoever. We show that this view is overly narrow. Many randomization tests are also valid under a more general "bounded" null hypothesis under which all effects are weakly negative (or positive), thus accommodating heterogenous effects. By inverting such tests we can form one-sided confidence intervals for the maximum (or minimum) effect. These properties hold for all effect-increasing test statistics, which include both common statistics such as the mean difference and uncommon ones such as Stephenson rank statistics. The latter's sensitivity to extreme effects permits detection of positive effects even when the average effect is negative. We argue that bounded nulls are often of substantive or theoretical interest, and illustrate with two applications: testing monotonicity in an IV analysis and inferring effect sizes in a small randomized experiment.

✉ luke_miratrix@gse.harvard.edu

RERANDOMIZATION AND ANCOVA

Peng Ding*, *University of California, Berkeley*

Covariates are important for the design and analysis of experiments. With a discrete covariate, blocking is used in design to improve the quality of experiments and post-stratification is used in the analysis to improve the efficiency of estimation. Extending the dual relationship between blocking and post-stratification, we show that rerandomization in design and ANCOVA in the analysis are the duals for general covariates in experiments. Our theory is purely randomization-based, without assuming any parametric models for the outcomes. (Joint work with Xinran Li).

✉ pengdingpku@gmail.com

III. FRONTIERS IN HIGH-DIMENSIONAL DATA & BIG DATA ANALYSIS

MEASUREMENT ERROR—NOT JUST A NUISANCE

Len Stefanski*, *North Carolina State University*

The talk will describe some creative uses of measurement error modeling concepts for the purpose motivating certain approaches to variable selection. The development from concept to method will be traced and methods so derived will be presented.

✉ stefansk@ncsu.edu

HYPOTHESIS TESTING ON LINEAR STRUCTURES OF HIGH DIMENSIONAL COVARIANCE MATRIX

Runze Li*, *The Pennsylvania State University*

Shurong Zheng, *Northeast Normal University*

Zhao Chen, *The Pennsylvania State University*

Hengjian Cui, *Capital Normal University*

We study test of significance on high dimensional covariance structures, and develop a unified framework for testing commonly-used linear covariance structures. We first construct a consistent estimator for parameters involved in the linear covariance structure, and then develop two tests for the linear covariance structures based on entropy loss and quadratic loss used for covariance matrix estimation. To study properties of the proposed tests, we study related high dimensional random matrix theory, and establish several highly useful asymptotic results. With the aid of these asymptotic results, we derive the limiting distributions of these two tests under the null and alternative hypotheses. We further show that the quadratic loss based test is asymptotically unbiased. We conduct Monte Carlo simulation study to examine the finite sample performance of the two tests. Our numerical comparison implies that the proposed tests outperform existing ones in terms of controlling Type I error rate and power. Our simulation indicates that the test based on quadratic loss seems to have better power than the test based on entropy loss.

✉ rzli@psu.edu

LARGE-SCALE INFERENCE WITH GRAPHICAL NONLINEAR KNOCKOFFS

Yingying Fan*, *University of Southern California*

Emre Demirkaya, *University of Southern California*

Gaorong Li, *Beijing University of Technology*

Jinchi Lv, *University of Southern California*

We provide theoretical foundations on the power and robustness for the modelfree knockoffs procedure introduced recently in Cand'es, Fan, Janson and Lv (2016) in high-dimensional setting when the covariate distribution is characterized by Gaussian graphical model. We establish that under mild regularity conditions, the power of the oracle knockoffs procedure with known covariate distribution in high-dimensional linear models is asymptotically one as sample size goes to infinity. When moving away from the ideal case, we suggest the modified model-free knockoffs method called graphical nonlinear knockoffs (RANK) to accommodate the unknown covariate distribution. We provide theoretical justifications on the robustness of our modified procedure by showing that the false discovery rate (FDR) is asymptotically controlled at the target level and the power is asymptotically one with the estimated covariate distribution. To the best of our knowledge, this is the first formal theoretical result on the power for the knockoffs procedure. Numerical results demonstrate that compared to existing approaches, our method performs competitively in both FDR control and power.

✉ fanyingy@marshall.usc.edu

NETWORK MEMBERSHIP ESTIMATION BY MIXED-SCORE

Tracy Ke*, *University of Chicago*

Jiashun Jin, *Carnegie Mellon University*

Shengming Luo, *Carnegie Mellon University*

Consider an undirected mixed membership network with n nodes and K communities. For each node i, we model the membership as a probability mass function a_i that takes K different values. We call node i pure if a_i is degenerate and mixed otherwise. The main interest is to esitmate a_i, i = 1,2, ..., n. We model the adjacency matrid of the network by a Degree-Corrected Network model, and reveal a surprising simplex structure associated with the first K eigenvectors of the adjacency matrix. The finding motivates us a new approach to membership estimation which we call Mixed-SCORE. We apply Mixed-SCORE to 4 different data sets with encouraging results. We explain the simplex structure is real and not coincident and show that Mixed-SCORE is optimal in membership estimation.

✉ jiashun@stat.cmu.edu

II2. FUNCTIONAL CONNECTIVITY AND NETWORKS

LATENT SOURCE SEPARATION FOR MULTI-SUBJECT BRAIN NETWORKS

Ben Wu*, *Emory University*

Jian Kang, *University of Michigan*

Ying Guo, *Emory University*

Advanced imaging techniques for human brain enable us to more accurately measure the structural connectivity and functional connectivity of anatomical regions for each subject, generating large-scale brain networks individually. It is very challenging but extremely important to jointly analyze brain networks of multiple subjects in order to have a better understanding of how the functioning of the human brain depends on its network architecture. In particular, it is of great interest to identify the important factors that explain the commonality and variations of brain networks across subjects. To this end, we propose a new latent source separation approach to decomposing multiple brain networks, where each subject-specific network is modeled as a mixture of latent source sub-networks. We develop an efficient computational estimation method. We examine the theoretical properties of the proposed model and the estimation procedure. We demonstrate the advantages of the proposed method compared with existing alternatives via simulation studies and data analysis for a neuroimaging study.

✉ ben.wu@emory.edu

A SPATIAL-TEMPORAL MODEL FOR DETECTING THE EFFECT OF COCAINE DEPENDENCE ON BRAIN CONNECTIVITY

Jifang Zhao*, *Virginia Commonwealth University*

Montserrat Fuentes, *Virginia Commonwealth University*

Liangsuo Ma, *Virginia Commonwealth University*

Frederick Gerard Moeller, *Virginia Commonwealth University*

Qiong Zhang, *Virginia Commonwealth University*

Brain Connectivity obtained from resting state data promotes a variety of fundamental understandings in Neuroscience. It is of particular interest in quantifying the treatment effects on brain connectivity. In this work, we propose a spatial-temporal model for multi-subjects resting state data. Our model characterizes the brain connectivities through a model covariance matrix. A hypothesis testing approach is proposed to quantify the significance of treatment effects on brain connectivity. Based on case/control resting state dataset collected from multiple subjects with or without cocaine dependence, our methods are applied to detecting the effect of cocaine dependence on brain connectivity.

✉ jifangzhao64@hotmail.com

EVIDENCE-BASED INFERENCE ON RESTING STATE FUNCTIONAL CONNECTIVITY

Allison E. Hainline*, *Vanderbilt University*

Hakmook Kang, *Vanderbilt University*

Traditional inferential methods pre-specify the Type I error and maximize the power, though with a large number of comparisons (as is necessary for several ROI-level functional connectivity, e.g., 45 comparisons for 10 ROIs), the global Type I error rate inflates rapidly. Applying such methods to identify significant functional connectivity within the brain can result in high Type I or Type II errors. Inflated global Type I error can be prevented via the likelihood paradigm, which uses the likelihood ratio as a measure of the strength of evidence, rather than traditional significance tests. A key result of the likelihood paradigm is the convergence of both global Type I and Type II error analogs to zero as the sample size increases, whereas traditional methods may never reach a Type I error below the pre-specified size of the test. We present an approach for identifying regions where the evidence of functional connectivity is strong, without the risk of inflated error rates. A simulation study shows the superiority of the likelihood-based method in terms of average error rate for small samples. We also provide an application of the methods to a study of 6 healthy subjects.

✉ allison.e.hainline@vanderbilt.edu

LATENT CLASSES OF RESTING-STATE FUNCTIONAL CONNECTIVITY AND THEIR ASSOCIATION WITH CLINICAL FEATURES

Xin Ma*, *Emory University*

Ying Guo, *Emory University*

Limin Peng, *Emory University*

Amita Manatunga, *Emory University*

Resting-state functional connectivity (rs-fc) derived from functional magnetic resonance imaging (fMRI) data offers insights on the organization of intrinsic brain functional networks. In this work, we investigate the latent classes of subject's rs-fc profiles and their association with clinical outcomes. Specifically, we extract latent rs-fc patterns from fMRI across subjects using the independent component analysis (ICA) and subject-specific loadings on these latent rs-fc patterns are then associated with subjects' clinical outcomes. We also investigate several clustering methods to identify homogeneous clusters based on rs-fc and their relationship with clinical features. We report our findings from the Grady Trauma Project.

✉ xin.ma@emory.edu

SPECTRAL DOMAIN COPULA-BASED DEPENDENCE MODELS WITH APPLICATION TO BRAIN SIGNALS

Charles Fontaine*, *King Abdullah University of Science and Technology*

Hernando Ombao, *King Abdullah University of Science and Technology*

The neuroscience community is currently putting a strong emphasis on studying brain connectivity. This is non-trivial since there is no unique measure that can adequately explain how brain regions communicate. In response to it, we will develop a more sophisticated dependence measure that can identify mechanisms of connectivity that are overlooked by standard measures such as correlation/coherence, full or partial. We propose a copula-based model that keeps the robustness of parametric models through the expression of a dependence parameter. In order to capture the dependence between brain regions through the oscillations (spectral), we work with margins of the time series through functionals of their Fourier transforms. Our proposed method expresses the multivariate copulas into a network of bivariate copulas. We propose a methodology to write multivariate copulas composed of the information of 32 brain channels into a network of bivariate copulas. From this network, we can illustrate dependence between the time series. We demonstrate this new method to a local field potential (LFP) data to study connectivity in the rat brain prior to and following an induced stroke.

✉ charles.fontaine@kaust.edu.sa

MODELING DYNAMIC BRAIN CONNECTIVITY

Marco Antonio Pinto Orellana*, *King Abdullah University of Science and Technology*

Ting Chee-Ming, *King Abdullah University of Science and Technology*

Jeremy Skipper, *University College London*

Steven Small, *University of California, Irvine*

Hernando Ombao, *King Abdullah University of Science and Technology*

It is widely accepted that connectivity between brain regions is dynamic. To characterize dynamic connectivity,we propose a method that consists of state identification,change-points estimation,and connectivity estimation. The first step is to recognize the states by detecting some K-means clusters,of stable connectivity,in time-varying vector autoregressive (TV-VAR) coefficients. The next task is to determine the regime change-points using the Kalman filter and smoother. Finally,the third stage is to estimate the network connectivity applying a switching VAR and an E-M algorithm. To demonstrate the utility of the proposed SVAR model,we analyze a fMRI data recorded while participants watched a movie. This is a complex experimental setting where the goals are to determine switches between brain states with respect to a number of communication markers such as the use of hand gestures and voice intonation. Our results show this method estimates steady and reproducible states in single-subject analyses. In addition,we developed a Python toolbox with a graphical user interface that allows for a seamless integration between visualization of results and model estimation and computation.

✉ pinto.marco@live.com

A NOVEL HIERARCHICAL INDEPENDENT COMPONENT MODELING FRAMEWORK WITH APPLICATION TO LONGITUDINAL fMRI STUDY

Yikai Wang* •, *Emory University*

Ying Guo, *Emory University*

Recently longitudinal neuroimaging study has become increasingly popular to study brain functional networks(BFN) related to clinical/demographical status, where the most commonly used tool is independent component analysis(ICA). However, existing ICA methods are only for cross-section study and not suited for repeatedly measured imaging data. Here, we propose a novel longitudinal ICA model(L-ICA) as the first formal statistical modeling framework to extend ICA to longitudinal study. By incorporating subject-specific random effect and visit-specific covariate effect, L-ICA is able to provide more accurate decomposition, borrow information within the same subject and provide model-based prediction. We develop traceable EM algorithm and further develop two approximate EM algorithms based on voxel-specific subspace and sub-sampling technique which greatly reduce computation time while retaining high accuracy. We also propose an approximate inference procedure for time-specific covariate effect. Simulation results demonstrate the advantages of our methods. We apply L-ICA to ADNI2 study and discover biologically insightful findings which are not revealed by existing methods.

✉ Yikai.wang@emory.edu

II3. INDIVIDUALIZED TREATMENT RULES

HYPOTHESIS TESTINGS ON INDIVIDUALIZED TREATMENT RULES FROM HIGH-DIMENSIONAL OBSERVATIONAL STUDIES

Young-Geun Choi*, *Fred Hutchinson Cancer Research Center*

Yang Ning, *Cornell University*

Yingqi Zhao, *Fred Hutchinson Cancer Research Center*

Individualized treatment rules (ITR) assign treatments according to different patient's characteristics. Despite recent advances on the estimation of ITRs, much less attention has been given to uncertainty assessments for the estimated rules. We propose a hypothesis testing procedure

for the estimated ITRs from a general framework that directly optimizes overall treatment benefit. Specifically, we construct a local test for testing low dimensional components of high-dimensional linear decision rules. The procedure can apply to observational studies by taking into account the additional variability from the estimation of propensity score. Theoretically, our test extends the decorrelated score test proposed in Nang and Liu (2017) and is valid no matter whether model selection consistency for the true parameters holds or not. The proposed methodology is illustrated with numerical studies and a real data example on electronic health records of patients with Type-II Diabetes.

✉ ychoi2@fredhutch.org

MATCHED LEARNING (M-LEARNING) FOR ESTIMATING OPTIMAL INDIVIDUALIZED TREATMENT RULES WITH AN APPLICATION TO ELECTRONIC HEALTH RECORDS

Peng Wu* •, *Columbia University*

Donglin Zeng, *University of North Carolina, Chapel Hill*

Yuanjia Wang, *Columbia University*

Individualized treatment rules (ITRs) tailor medical treatments according to individual-specific characteristics.It is of interest to learn effective ITRs using data collected in practical clinical settings. In this work, we propose a machine learning approach to estimate ITRs through matching methods, referred as M-learning. M-learning is applicable to observational studies as well as randomized controlled trials (RCTs). It performs matching instead of inverse probability weighting to more accurately estimate an individual treatment response under alternative treatments and alleviate confounding in observational studies. A matching function is proposed to compare outcomes for matched pairs where various types of outcomes can be easily accommodated. We further improve efficiency of estimating ITR by a denoise procedure and double robust matching. We prove Fisher consistency of M-learning and conduct extensive simulation studies. We show that M-learning outperforms some existing methods when propensity scores are misspecified and in certain scenarios of the presence of unmeasured confounders. Lastly, we apply our method to a study using electronic health records.

✉ pw2394@cumc.columbia.edu

SEQUENTIAL OUTCOME-WEIGHTED MULTICATEGORY LEARNING FOR ESTIMATING OPTIMAL INDIVIDUALIZED TREATMENT RULES

Xuan Zhou*, *University of North Carolina, Chapel Hill*

Yuanjia Wang, *Columbia University*

Donglin Zeng, *University of North Carolina, Chapel Hill*

Powerful machine learning methods have been proposed to estimate an optimal individualized treatment rule (ITR), but they are developed for binary treatment decisions and thus limited to compare only two treatments. When many treatment options are available, existing methods need to be adapted by transforming a multicategory treatment selection problem into multiple binary ones. However, how to combine multiple binary treatment selection rules into a single decision rule is not straightforward and it is well known in the multicategory learning literature that some approaches may lead to inconsistent decision rules. In this article, we propose a novel and efficient method to generalize outcome weighted learning to multi-treatment settings. Specifically, we solve a multicategory treatment selection problem via sequential weighted support vector machines. Theoretically, we show that the resulting ITR is Fisher consistent. We demonstrate the performance of the proposed method with extensive simulations and an application to a three-arm randomized trial of treating major depressive disorder.

✉ xuanz@live.unc.edu

SEMIPARAMETRIC SINGLE-INDEX MODELS FOR OPTIMAL TREATMENT REGIMES WITH POTENTIALLY CENSORED OUTCOMES

Jin Wang*, *University of North Carolina, Chapel Hill*

Donglin Zeng, *University of North Carolina, Chapel Hill*

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

We propose a class of proportional hazards models with single-index functions to model interactions between treatment and covariates. The new models are exible enough to incorporate non-linear treatment-covariate interactions but still provide a simple and interpretable optimal treatment rule. We develop an estimation procedure for the maximum likelihood estimators and establish the asymptotic properties for the estimators. We conduct simulation studies to evaluate the nite sample performance of the proposed method. We apply the proposed model to the AIDS Clinical Trial Group study to illustrate how treatment effect varies across patients depending on baseline variables, and the estimation of individual treatment regimes in practical settings.

✉ jinjin@live.unc.edu

LASSO FOR MODELING TREATMENT COVARIATE INTERACTIONS

Yu Du*, *Johns Hopkins University*

Ravi Varadhan, *Johns Hopkins School of Medicine*

We propose a general method to predict treatment response heterogeneity, through identification of treatment covariate interactions. We construct a single-step L1 penalty procedure that maintains the interaction hierarchy in a sense that the interaction term is included in the model only when its associated main effect term is included. We explore several parameterization schemes that enforce the interaction hierarchy using Lasso, and we solve the constrained optimization problem using spectral projected gradient method. Extensive simulation studies are conducted, considering a wide variety of scenarios for treatment covariate interactions. The study shows that our method results in a more parsimonious model than regular Lasso, and produces better performance than the regular Lasso and traditional two-step regression with backward selection in terms of prediction performance. We also apply our method to a large randomized clinical trial data on a drug for treating congestive heart failure. Our method provides an attractive alternative, with sufficient flexibility in terms of parametrization, to model individualized treatment effect with interaction hierarchy.

✉ ydu10@jhu.edu

II4. METABOLOMICS AND PROTEOMICS

BAYESIAN LATENT CLASS MODELS FOR IDENTIFYING CIRCADIAN PATTERNS IN HIGH-DIMENSIONAL LONGITUDINAL METABOLOMICS DATA

Sung Duk Kim*, *National Cancer Institute, National Institutes of Health*

Paul S. Albert, *National Cancer Institute, National Institutes of Health*

Many researchers in biology and medicine have focused on trying to understand circadian rhythms and their potential impact on disease. A recent pilot study measured metabolomic profiles at various times of the day in a group of health volunteers who were enrolled in an inpatient circadian rhythm laboratory without day or night cues. Of interest was identifying the number and type of metabolites that showed circadian patterns in this pristine environment. We were also interested in whether the same metabolites showed similar phase shifts across subjects.We develop new statistical methodology for identifying biomarkers that exhibit circadian patterns among these high-dimensional metabolomic profiles. A latent class approach is proposed to separate individual metabolomic profiles into groups with differing circadian patterns and a group with no such pattern. Markov chain Monte Carlo sampling is used to carry out Bayesian posterior computation. Several variations of the proposed models are considered and compared using the deviance information criterion. The model is used to reveal important insight into the circadian biology in metabolomics.

✉ kims2@mail.nih.gov

BAYESIAN NONPARAMETRIC MATCHING OF CELL LINES TO TUMOR SAMPLES

Chiyu Gu*, *University of Missouri*

Subharup Guha, *University of Missouri*

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

Immortalized cancer cell lines, derived from tumors and grown and maintained in vitro, are commonly used in the study of cancer biology. Although cancer cell lines are expected to closely resemble the particular cancer type of interest, various studies have identified molecular differences between commonly used cancer cell lines and tumor samples. With the protein expression data across a large number of cancer cell lines recently available through reverse-phase protein arrays, we are now able to thoroughly evaluate the similarity of protein expression between cell lines and tumor samples. We propose an innovative method to compare cell lines to tumor samples using proteomics data based on Bayesian nonparametric models. The proposed method introduces a flexible, bi-directional model based clustering mechanism that simultaneously cluster the protein markers and samples. The proposed method not only provides better accuracies, but also more detailed quantification of similarities compared to traditional clustering methods. We demonstrate the effectiveness of the proposed method by the analysis on the proteomics data from The Cancer Proteome Atlas (TCPA) for lung cancer.

WITHDRAW

✉ cgz59@mail.missouri.edu

MISSING DATA IMPUTATION FOR MASS SPECTROMETRY METABOLOMICS DATA USING A 2-STEP LASSO APPROACH

Qian Li*, *University of South Florida*

Brooke L. Fridley, *Moffitt Cancer Center*

Chengpeng Bi, *Children's Mercy Hospital*

Roger Gaedigk, *Children's Mercy Hospital*

Steven Leeder, *Children's Mercy Hospital*

Imputation is one of the challenges in analyzing mass spectrometry (MS) metabolites data. Existing research on metabolomics imputation suggests a modified KNN method, i.e. KNN truncation to account for non-random missing, which still cannot improve prediction for missing at random. We proposed a 2-step LASSO imputation method: first predict missing values with metabolites without missing values by LASSO, and then repeat such procedure on the imputed data to address both types of missing simultaneously. This method is compared with its first step and KNN truncation, using 98 liquid chromatography (LC)-MS metabolite samples from a pediatric liver study conducted at Children's Mercy Hospital to look at changes in metabolite levels over childhood development. We randomly 'masked' samples for each metabolite, then imputed masked data by each method. Performance is evaluated by correlation between predicted and true values for each metabolite, showing that the 2-step LASSO outperforms KNN truncation with median Pearson correlation increased from 0.75 to 0.85.

✉ qian.li10000@gmail.com

A TWO-PART SEMI-PARAMETRIC MODEL FOR ZERO-INFLATED METABOLOMICS AND PROTEOMICS DATA

Yuntong Li*, *University of Kentucky*

Identifying differentially abundant features between different experimental conditions is a common goal for many metabolomics and proteomics studies. However, analyzing metabolomics and proteomics data from mass spectrometry is challenging because the data may not be normally distributed and contain large fraction of zero values. Although several statistical methods have been proposed, they either require data normality assumption, or are inefficient. We propose a new Semi-parametric Differential Abundance analysis (SDA) method for metabolomics and proteomics data from mass spectrometry.The method considers a

two-part model, a logistic regression for the zero proportion and a semi-parametric log-linear model for the non-zero values. Our method is free of distributional assumption and also allows for adjustment of covariates. We propose a kernel-smoothed likelihood method to estimate regression coefficients in the two-part model and construct a likelihood ratio test for differential abundant analysis. Simulations and real data analyses demonstrate that our method outperforms existing methods.

✉ yli362@g.uky.edu

II5. BAYESIAN HIGH DIMENSIONAL DATA AND VARIABLE SELECTION

HIGH DIMENSIONAL POSTERIOR CONSISTENCY IN BAYESIAN VECTOR AUTOREGRESSIVE MODELS

Satyajit Ghosh*, *University of Florida*

Kshitij Khare, *University of Florida*

George Michailidis, *University of Florida*

Vector autoregressive (VAR) models aim to capture linear temporal interdependencies among multiple time series. They have been widely used in macro and financial econometrics and more recently have found novel applications in functional genomics and neuroscience. These applications have also accentuated the need to investigate the behavior of the VAR model in a high-dimensional regime, which will provide novel insights into the role of temporal dependence for regularized estimates of the model's parameters. However, hardly anything is known regarding properties of the posterior distribution for Bayesian VAR models in such regimes. In this work, we consider a VAR model with two prior choices for the autoregressive coefficient matrix: a non-hierarchical matrix-normal prior and a hierarchical prior which corresponds to an arbitrary scale mixture of normals. We establish posterior consistency for both these priors under standard regularity assumptions, when the dimension p of the VAR model grows with the sample size n (but still remains smaller than n).

✉ satyajitghosh90@ufl.edu

MORETreeS: A FLEXIBLE METHOD FOR MULTI-OUTCOME REGRESSION WITH TREE-STRUCTURED SHRINKAGE

Emma G. Thomas*, *Harvard University*

Giovanni Parmigiani, *Harvard University*

Francesca Dominici, *Harvard University*

Lorenzo Trippa, *Harvard University*

We present Multi-Outcome Regression Estimation with Tree-structured Shrinkage (MORETreeS), a novel class of Bayesian regression models for use in estimating the effect of a common exposure on multiple outcomes whose relationships follow a tree structure. MORETreeS employs a general class of tree-structured priors that enable shrinkage of the parameters for related outcomes towards one another with the potential for fusion. We focus on count data, for which the model reduces the effective number of outcomes by automatically summing over related outcomes when effects appear homogeneous or counts are small, whilst estimating separate coefficients when there is sufficient evidence of exposure effect heterogeneity. As a motivating example, we consider the effect of PM2.5 exposure on hospitalizations tagged with International Classification of Diseases codes. MORETreeS is most useful when the number of outcomes is greater than the number of observations and the number of independent variables is small. We recommend use of MORETreeS when little or no confounding adjustment is required or in conjunction with two-stage causal inference methods such as propensity score matching.

✉ emmathomas@g.harvard.edu

HIGH-DIMENSIONAL POSTERIOR CONSISTENCY FOR HIERARCHICAL NON-LOCAL PRIORS IN REGRESSION

Xuan Cao*, *University of Florida*

Malay Ghosh, *University of Florida*

Kshitij Khare, *University of Florida*

The choice of tuning parameters in Bayesian variable selection is a critical problem in modern statistics. Especially in the related work of non-local prior under regression setting, the scale parameter reflects the dispersion of the non-local prior density around zero, and implicitly determines the size of the regression coefficients that will be shrunk to zero. In this paper, we introduce a fully Bayesian approach with the pMOM non-local prior where we place an appropriate Inverse-Gamma prior on the tuning parameter to analyze a more robust model that is comparatively immune to misspecification of scale parameter. Under standard regularity assumptions, we extend the previous work where p is bounded by the number of observations n and establish strong model selection consistency when p is allowed to increase at a polynomial rate with n. Through simulation studies, we demonstrate that our model selection procedure outperforms commonly used penalized likelihood methods and other Bayesian methods with fixed parameters in a range of simulation settings.

✉ njualicia@gmail.com

FROM MIXED-EFFECTS MODELING TO SPIKE AND SLAB VARIABLE SELECTION: A BAYESIAN REGRESSION MODEL FOR GROUP TESTING DATA

Chase Joyner*, *Clemson University*

Christopher McMahan, *Clemson University*

Joshua Tebbs, *University of South Carolina*

Christopher Bilder, *University of Nebraska-Lincoln*

Due to reductions in cost, group testing is becoming a popular alternative to individual level testing. These reductions are gained by testing pooled bio-specimen (e.g., blood) for the presence of an infectious agent. Though this process may reduce cost, it comes at the expense of data complexity. Further, in venues in which group testing is employed, practitioners are faced with conducting surveillance. To this end, one desires to fit a regression model which relates individual level covariate information to disease status. This is a nontrivial task since an individual's disease status is obscured by imperfect testing. Further, unlike individual level testing, a given participant could be involved in multiple testing outcomes. To circumvent these hurdles, we propose a Bayesian generalized linear mixed model which can accommodate data arising from any group testing procedure and can account for heterogeneity in the covariate effects across clinic sites. This proposal makes use of spike and slab priors, for the fixed and random effects, to achieve model selection. For illustrative purposes, the approach is applied to a Chlamydia data collected in Iowa.

✉ chasej@clemson.edu

ROBUST BAYESIAN VARIABLE SELECTION FOR MODELING MEAN MEDICAL COSTS

Grace Yoon*, *Texas A&M University*

Wenxin Jiang, *Northwestern University*

Lei Liu, *Washington University in St. Louis*

Ya-Chen Tina Shih, *University of Texas MD Anderson Cancer Center*

Several statistical issues associated with health care costs, such as heteroscedasticity and severe skewness, make it challenging to estimate medical costs. When modeling the mean cost, it is desirable to make no assumption on the density function or higher order moments. Another challenge in developing cost prediction model is the presence of many covariates. Variable selection is needed to achieve a balance of prediction accuracy and model simplicity. We propose Spike-or-Slab priors for Bayesian variable selection based on asymptotic normal estimates of the full model

parameters that are consistent as long as the assumption on the mean cost is satisfied. This method possesses three advantages: robustness (due to avoiding assumptions on the density function or higher order moments), parsimony (feature of variable selection), and expressiveness (due to its Bayesian flavor, which can compare posterior probabilities of candidate models). In addition, by ranking the Z-statistics, the scope of model searching can be reduced to achieve computational efficiency. We apply this method to the Medical Expenditure Panel Survey dataset.

✉ gyoon@stat.tamu.edu

BAYESIAN VARIABLE SELECTION FOR MULTI-OUTCOME MODELS THROUGH SHARED SHRINKAGE

Debamita Kundu*, *University of Louisville*

Jeremy Gaskins, *University of Louisville*

Ritendranath Mitra, *University of Louisville*

In Bayesian context, variable selection over a potentially large set of covariates of a linear model is quite popular. The common prior choices lead to a posterior expectation of the regression coefficients that is a sparse (or nearly sparse) vector with a few non-zero components, those covariates that are most important. This project is motivated by the "global-local" shrinkage prior idea. Here we have developed a variable selection method for a K-outcome model that identifies the most important covariates across all outcomes. We consider two versions of our approach based on the normal-gamma prior (Griffin & Brown, 2010, Bayesian Analysis) and the Dirichlet-Laplace prior (Bhattacharya et al., 2015, JASA). The prior for all regression coefficients is a mean zero normal with coefficient-specific variance term that consists of a predictor-specific factor and a model-specific factor. The predictor-specific terms play that role of a shared local shrinkage parameter, whereas the model-specific factor is similar to a global shrinkage term that differs in each model. The performance of our modeling approach is evaluated through a simulation study and data example.

✉ debamita.kundu@louisville.edu

II6. METHODS FOR SURVIVAL ANALYSIS

EMPIRICAL COMPARISON OF THE BRESLOW ESTIMATOR AND THE KALBFLEISCH-PRENTICE ESTIMATOR FOR SURVIVAL FUNCTIONS

Fang Xia*, *University of Texas MD Anderson Cancer Center*

Xuelin Huang, *University of Texas MD Anderson Cancer Center*

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

Jack D. Kalbfleisch, *University of Michigan*

The Cox proportional hazards model is commonly used to analyze time-to-event data. To estimate the survival function, both the Breslow estimator and the Kalbfleisch-Prentice (K-P) estimator can be used. The Breslow estimator will never reach zero since it is an exponential function of the cumulative baseline hazards. On the other hand, the K-P estimator is similar to the Kaplan-Meier estimator. It will reach zero if the last observation is an event. It accounts for covariate effects through the Cox model. In order to evaluate the relative performance of these estimators, we conducted simulation studies across a range of conditions. We varied the true survival time distribution, sample size, censoring rate and covariate values, and compared the performance according to bias, mean square error and relative mean square error. In most situations of our study, the K-P estimator outperformed the Breslow estimator on estimating survival probabilities with less bias and smaller mean square error. Their differences are especially clear at the tail. This suggests that the K-P estimator may provide a more accurate estimation of cure rate than the Breslow estimator.

✉ fang.xia@uth.tmc.edu

TIME-VARYING PROPORTIONAL ODDS MODEL FOR MEGA-ANALYSIS OF CLUSTERED EVENT TIMES

Tanya Garcia*, *Texas A&M University*

Karen Marder, *Columbia University*

Yuanjia Wang, *Columbia University*

Mega-analysis, or the meta-analysis of individual data, enables pooling and comparing multiple studies to enhance estimation and power. A challenge in mega-analysis is estimating the distribution for clustered, potentially censored event times where the dependency structure can introduce bias if ignored. We propose a new proportional odds model with unknown, time-varying coefficients and multi-level random effects. The model directly captures event dependencies, handles censoring using pseudo-values, and permits a simple estimation by transforming the model into an easily estimable additive logistic mixed effect model. Our method consistently estimates the distribution for clustered event times even under covariate-dependent censoring. Applied to three observational studies of Huntington's disease, our method provides, for the first time in the literature, evidence of similar conclusions about motor and cognitive impairments in all studies despite different recruitment criteria.

✉ tpgarcia@sph.tamhsc.edu

COX REGRESSION FOR RIGHT-TRUNCATED DATA

Bella Vakulenko-Lagun*, *Harvard University*

Micha Mandel, *Hebrew University of Jerusalem*

Rebecca A. Betensky, *Harvard University*

Right-truncated survival data arise when observations are ascertained retrospectively and only those who experienced the event of interest prior to sampling are retained in the study. When the main interest is in the association between the time to event and covariates, a Cox proportional hazards model is often used for analysis. However, for right-truncated data there is no such likelihood factorization that yields an estimating equation involving only covariate effects. We consider two methods for handling right-truncated data. One method involves "reversing" time - it can be used for inference about the direction of association in forward time and testing both global and partial hypotheses about forward-time covariate effects but not for their estimation. The second method is based on Inverse-Probability-Weighting estimating equations, which do allow estimation of forward-time covariate effects under a positivity assumption. We discuss the problems of identifiability and consistency that might arise with right-truncated data and compare these two methods to other approaches.

✉ blagun@hsph.harvard.edu

MODELING NEGATIVELY SKEWED SURVIVAL DATA IN AFT AND CORRELATED FRAILTY MODELS USING THE RSTG DISTRIBUTION

Sophia D. Waymyers*, *Francis Marion University*

Hrishikesh Chakraborty, *Duke Clinical Research Institute*

Negatively skewed survival data occasionally arises in medical research. The reflected-shifted-truncated gamma (RSTG) distribution has been shown to be effective in modeling negatively skewed survival data. We examine the efficacy of the RSTG distribution in the accelerated failure time (AFT) model and for the correlated gamma frailty model in both simulated negatively skewed data and data from the 1972 Diabetic Retinopathy study. Maximum likelihood methods are used for parameter estimation. We use information theoretic criteria to show that the RSTG AFT model performs better than the AFT model with exponential, generalized F, generalized gamma, Gompertz, log-logistic, lognormal, Rayleigh or Weibull distributional assumptions. The effectiveness of the RSTG distribution in the correlated gamma frailty model applied to the diabetic retinopathy study is also illustrated, and deviance residual plots reveal that the model is a good fit. The RSTG distribution performs well in both the AFT model and for the correlated gamma frailty model when modeling negatively skewed survival data.

✉ swaymyers@fmarion.edu

INTEGRATIVE SURVIVAL ANALYSIS WITH UNCERTAIN EVENT RECORDS WITH APPLICATION IN A SUICIDE RISK STUDY

Wenjie Wang*, *University of Connecticut*

Robert Aseltine, *UConn Health*

Kun Chen, *University of Connecticut*

Jun Yan, *University of Connecticut*

The concept of integrating data from disparate sources to accelerate scientific discovery has generated tremendous excitement in many fields. The potential benefits from data integration, however, may be compromised by the uncertainty due to imperfect record linkage. Motivated by a suicide risk study, we propose an approach for analyzing survival data with uncertain event records arising from data integration. We develop an integrative Cox proportional hazards model, in which the uncertainty in the matched event times is modeled probabilistically. The estimation procedure combines the ideas of profile likelihood and the expectation conditional maximization algorithm. Simulation studies demonstrate that under realistic settings, the proposed method outperforms several competing approaches including multiple imputation. A marginal screening analysis using the proposed integrative Cox model is performed to identify risk factors associated with death following suicide-related hospitalization in Connecticut. The identified diagnostics codes are consistent with existing literature and provide several new insights on suicide risk prediction and prevention.

✉ wenjie.2.wang@gmail.com

NEW MODELING OF RECURRENT EVENTS DATA SUBJECT TO TERMINAL EVENTS

Bo Wei*, *Emory University*

Limin Peng, *Emory University*

Zhumin Zhang, *University of Wisconsin, Madison*

Huichuan Lai, *University of Wisconsin, Madison*

Recurrent events data are commonly encountered in biomedical follow-up studies. Often the observation of recurrent events can be terminated by a terminal event such as death or drop-off. In this work, we investigate how to apply or adapt the generalized accelerated recurrence time (GART) model to recurrent events data subject to terminal events. Our investigations lead to two different modeling that appropriate accounts for the presence of terminal events: one formulates covariate effects on the time to expected adjusted rate, while the other provides interpretation based on survivors' rate function. We provide justifications for both types of modeling, and develop estimation and inference accordingly. We conducted extensive simulation studies to evaluate the proposed approaches. We also illustrate our method via an application to a dataset from Cystic Fibrosis Foundation Patient Registry (CFFPR).

✉ bo.wei@emory.edu

OPTBAND: OPTIMAL CONFIDENCE BANDS FOR FUNCTIONS TO CHARACTERIZE TIME-TO-EVENT DISTRIBUTIONS

Sam Tracy*, *Harvard School of Public Health*

Tom Chen, *Harvard School of Public Health*

Hajime Uno, *Harvard Medical School*

Classical simultaneous confidence bands for survival functions (i.e. Hall-Wellner and Equal Precision) are derived from transformations of the asymptotic Brownian nature of the Nelson-Aalen or Kaplan-Meier estimators. Due to the properties of Brownian motion, a theoretical derivation of the highest confidence density region cannot be obtained in closed form. Instead, we provide confidence bands derived from a related optimization problem with local times. These bands can be applied for the one-sample problem regarding both cumulative hazard and survival functions, and the two-sample problem regarding cumulative hazard only. The finite-sample performance of the proposed method is assessed by Monte Carlo simulation studies. The proposed bands are applied to clinical trial data to assess the

cardiovascular efficacy and safety of saxaglipitin, a DPP-4 inhibitor, when added to standard of care in patients with type 2 diabetes mellitus.

✉ stracy@g.harvard.edu

II7. VARIABLE SUBSET AND MODEL SELECTION

WEAK SIGNALS IN HIGH-DIMENSION REGRESSION: DETECTION, ESTIMATION AND PREDICTION

Yanming Li*, *University of Michigan*

Hyokyoung Grace Hong, *Michigan State University*

S. Ejaz Ahmed, *Brock University, Canada*

Yi Li, *University of Michigan*

Regularization methods, including the Lasso, group Lasso and SCAD, typically focus on selecting variables with strong signals, while ignoring weak signals. This may result in biased prediction, especially when weak signals outnumber strong signals. In this work, we propose a covariance-insured variable selection method to detect weak signals that are partially correlated with strong signals and show that incorporating the weak signals can improve prediction. We further propose a post-selection shrinkage estimation procedure for weak signals. We establish asymptotic properties for the proposed method and evaluate its finite sample performance through simulation studies and a real data analysis.

✉ liyanmin@umich.edu

VARIABLE SELECTION AND PREDICTION IN TWO-PART REGRESSION MODELING FOR SEMICONTINUOUS DATA

Seongtae Kim*, *North Carolina A&T State University*

Semicontinuous data consist of a number of zero values and a distribution of positive values. This type of data is observed in many disparate application areas including health insurance expenditures, household saving and debt, and alcohol consumption. Semicontinuous data are often modeled using a two-part model. The first part models the probability of dichotomous outcomes, zero or positive, and the second part models the distribution of positive values. Despite the popularity of the two-part model, variable selection of the two-part model remains not fully addressed. Our objective is to investigate various variable selection techniques including information criterion methods and penalized methods in two-part models. Selection and prediction performances of selected techniques are evaluated via simulation studies, which consider zero-value inflation, multicollinearity, high dimensionality, and sparsity. Selection techniques are applied to community-based crime data where several violent crime response variables show semicontinuity, and some of over one hundred predictors possess multicollinearity.

✉ skim@ncat.edu

GAUSSIAN PROCESS SELECTIONS IN SEMIPARAMETRIC REGRESSION FOR MULTI-PATHWAY ANALYSIS

Jiali Lin*, *Virginia Tech*

Inyoung Kim, *Virginia Tech*

Analysis on clinical effects and genetic effects can have a huge impact in disease prediction. Particularly, it is important to identify significant genetic pathway effects associated to biomarkers. In this talk, we consider a problem of variable selection in a semiparametric regression model which can study the effects of clinical covariates and expression levels of multiple gene pathways. We model the unknown high-dimension functions of multi-pathways via Gaussian kernel machine to consider the possibility that genes within the same pathway interact with each other. Hence, our variable selection can be considered as Gaussian process selection which is associated to clinical outcome. We develop our Gaussian process selection under the Bayesian variable selection framework. We incorporate prior knowledge for structural pathways by imposing an Ising prior on the model. Our approach can be easily applied in high-

dimensional space where the sample size is smaller than the number of genes and covariates. We devise an efficient variational Bayes algorithm. Two simulations show us that our method has great power to catch significant pathways.

✉ jiali@vt.edu

VARIABLE SCREENING FOR HIGH DIMENSIONAL TIME SERIES

Kashif Yousuf*, *Columbia University*

Variable selection is a widely studied problem in high dimensional statistics, primarily since estimating the precise relationship between the covariates and the response is of great importance in many scientific disciplines. Given that high dimensional time series data sets are becoming increasingly common in many scientific disciplines such as neuroscience, public health and epidemiology, it is surprising that very few works on variable screening are applicable to time dependent data, or which attempt to utilize the unique structure of times series data. This paper introduces a generalized least squares screening (GLSS) procedure for high dimensional linear models with dependent and/or heavy tailed covariates and errors. As opposed to the original Sure independence screening (SIS) procedure, our attempts to utilizes the serial correlation present in the data when estimating our marginal effects. By utilizing this additional information GLSS is shown to outperform SIS in many cases.Sure screening properties are given for our procedure, and simulations are performed to demonstrate the finite sample performance of our procedure.

✉ ky2304@columbia.edu

EXAMINING THE VANISHING TETRAD NUMBER AS AN INDEX OF THE COMPLEXITY OF SEM MODELS

Hangcheng Liu*, *Virginia Commonwealth University*

Robert A. Perera, *Virginia Commonwealth University*

Structural Equation Modeling (SEM) is a series of statistical methods that allow complex relationships between one or more independent variables and one or more dependent variables, which is widely used in the behavioral sciences (e.g. psychology, psychobiology, sociology). Model complexity is defined as a model's average ability to fit different data patterns and it plays an important criteria to do model selection. Like linear regression, the number of free model parameters (q) is often used in traditional SEM model fit indices as a measure of the model complexity. However, only using q to indicate SEM model complexity is crude since other contributing factors, such as types of constraints or functional form are ignored. To solve this problem, we propose a special technique, Confirmatory Tetrad Analysis (CTA) to be a complement of traditional methods to test the model fit. The purpose of this study is to examine whether SEM model Complexity is related to the number of vanishing tetrads implied in the models.

✉ hangcheng1989@gmail.com

II8. METHODS FOR SINGLE-CELL ANALYSIS

ZINB-WaVE: A GENERAL AND FLEXIBLE METHOD FOR THE SUPERVISED AND UNSUPERVISED ANALYSIS OF SINGLE-CELL RNA-Seq

Davide Risso*, *Weill Cornell Medicine*

Fanny Perraudeau, *University of California, Berkeley*

Svetlana Gribkova, *Université Paris Diderot*

Sandrine Dudoit, *University of California, Berkeley*

Jean-Philippe Vert, *MINES ParisTech*

Single-cell RNA sequencing (scRNA-seq) is a powerful technique that enables researchers to measure gene expression at the resolution of single cells. Because of the low amount of RNA present in a single cell, many genes fail to be detected even though they are expressed; these genes are usually referred to as dropouts. Here, we present a general and flexible zero-inflated negative binomial model (ZINB-WaVE), which leads to low-dimensional representations of the data that account for zero inflation (dropouts),

over-dispersion, and the count nature of the data. We demonstrate, with simulations and real data, that the model and its associated estimation procedure are able to give a more stable and accurate low-dimensional representation of the data than PCA. Furthermore, the model can be used to compute cell-specific weights to unlock bulk RNA-seq DE pipelines for zero-inflated data.

✉ dar2062@med.cornell.edu

IDENTIFYING DIFFERENTIAL ALTERNATIVE SPLICING EVENTS USING SINGLE-CELL RNA SEQUENCING DATA

Yu Hu*, *University of Pennsylvania Perelman School of Medicine*

Nancy Zhang, *University of Pennsylvania*

Mingyao Li, *University of Pennsylvania Perelman School of Medicine*

The emergence of single-cell RNA-seq (scRNA-seq) technology has made it possible to measure gene expression at cellular level. This breakthrough enables a wider range of research studies such as exploring splicing heterogeneity among individual cells. However, compared to bulk RNA-seq, there are two unique challenges for scRNA-seq analysis: high technical variability and low sequencing depth. To overcome these challenges, we proposed a statistical framework, SCATS (Single-Cell Analysis of Transcript Splicing), which achieves high sensitivity at low coverage by accounting for technical noise. SCATS has two major advantages. First, it employs an empirical Bayes approach to model technical noise by use of external RNA spike-ins. Second, it groups "exons" originated from the same isoforms, which reduces the multiple testing burden and allows more informative reads to be utilized for detecting splicing change. We evaluate the performance of SCATS by extensive simulations and the analysis of real scRNA-seq data. We believe that it will improve the power in identifying differential alternative splicing events in scRNA-seq studies.

✉ huyu1@pennmedicine.upenn.edu

A BAYESIAN HIERARCHICAL MODEL FOR CLUSTERING DROPLET-BASED SINGLE CELL TRANSCRIPTOMIC DATA FROM MULTIPLE INDIVIDUALS

Zhe Sun*, *University of Pittsburgh*

Li Zhu, *University of Pittsburgh*

Ying Ding, *University of Pittsburgh*

Wei Chen, *Children's Hospital of Pittsburgh of UPMC, University of Pittsburgh*

Ming Hu, *Cleveland Clinic Foundation*

Single cell transcriptome sequencing (scRNA-Seq) has become a revolutionary tool. The newly developed droplet-based technologies enable efficient parallel processing of tens of thousands of single cells with direct counting of transcript copies using Unique Molecular Identifier (UMI). One challenge in the analysis of scRNA-Seq data is to identify cell subtypes from heterogeneous tissues. Multiple statistical approaches have been proposed to solve this problem. Among them, our previous work DIMM-SC can explicitly model UMI count data via a Dirichlet mixture model. However, most clustering methods (including DIMM-SC) only model variations among different single cells from a single individual. To jointly model data from multiple individuals, the individual level heterogeneity should be taken into consideration. To achieve this goal, we propose a novel Bayesian hierarchical model to simultaneously cluster scRNA-Seq data from multiple individuals. In both simulation studies and real data analysis, our proposed method outperforms other existing single-individual-based methods in terms of clustering accuracy and stability, and thus will facilitate novel biological discoveries.

✉ zhs31@pitt.edu

INTERROGATION OF HUMAN HEMATOPOIETIC TRAITS AT SINGLE-CELL AND SINGLE-VARIANT RESOLUTION

Caleb A. Lareau*, *Harvard University*

Two outstanding challenges in the post genome-wide association study (GWAS) era are 1) the precise identification of causal variants from implicated regions and 2) inference of pertinent cell types to heritable phenotypes. Here, we identify 36,919 unique fine-mapped variants with $>$ 1% causal posterior probability for a GWAS (n $\sim$113,000) for 16 blood cell traits. Though most putative causal variants fall outside known protein-coding regions, we observe significant enrichments of these loci in accessible chromatin of hematopoietic cell types. Pairing our putative causal variants with single-cell epigenomic measurements for $>$2,000 hematopoietic progenitor cells, we propose novel approaches for resolving the genetic architecture enrichments. We observe significant heterogeneity of trait enrichment within immunophenotypically-homogenous progenitor populations, notably common myeloid progenitors and megakaryocyte-erythroid progenitors, for several erythroid-related traits. In total, our mapping of hematopoietic traits at the single-cell and single-variant resolution provides a useful framework for dissecting associations generated by GWAS.

✉ caleblareau@g.harvard.edu

PENALIZED LATENT DIRICHLET ALLOCATION MODEL IN SINGLE CELL RNA SEQUENCING

Xiaotian Wu* •, *Brown University*

Zhijin Wu, *Brown University*

Hao Wu, *Emory University*

Single cell RNA sequencing (scRNA-seq) is a recently developed technology that allows quantification of RNA transcripts at individual cell level, providing cellular level resolution of gene expression variation. The scRNA-seq data are counts of RNA transcripts of all genes in species' genome. We adapt the Latent Dirichlet Allocation (LDA), a generative probabilistic model originated in natural language processing (NLP), to model the scRNA-seq data by considering genes as words and cells as documents, and latent biological functions as topics. In LDA, each documents is considered as the result of words generated from a mixture of topics, each with a different word usage frequency profile. We propose a penalized version of LDA to reflect the structure in scRNAseq, that only a small subset of genes are expected to be topic-specific. We apply the penalized LDA to two scRNA-seq data sets to illustrate the usefulness of the model. Using inferred topic frequency instead of word frequency substantially improves the accuracy in cell type classification.

✉ xiaotian_wu@brown.edu

GENE CO-EXPRESSION NETWORK ESTIMATION FROM SINGLE-CELL RNA-SEQUENCING DATA

Jihwan Oh*, *University of Pennsylvania*

Changgee Chang, *University of Pennsylvania*

Mingyao Li, *University of Pennsylvania*

Qi Long, *University of Pennsylvania*

The advent of single-cell RNA-sequencing (scRNA-seq) technology has enabled the study of gene expression properties, including gene-to-gene relationships, at the single-cell level. With scRNA-seq data, it is now possible to construct a high-resolution co-expression network at the subject level to better predict gene function, identify functional modules, and determine disease subtypes than previous bulk RNA-seq studies. However, scRNA-seq data are often noisy and in particular the large amount of zeros in gene expression makes previously developed network inference methods unsuitable for scRNA-seq data analysis. To tackle these challenges, we propose a new method that specifically accounts for technical noise and zero-inflation in scRNA-seq data. Our method employs a Bayesian hierarchical model with a multi-layered latent structure. This framework allows us to estimate the true correlation of gene expression for each gene pair, which in turn enables the construction of co-expression network for a set of genes. We evaluate the performance of our method by simulations and analysis of a murine brain single-cell dataset.

✉ jihwan05@gmail.com

I19. MODERN HIERARCHICAL APPROACHES TO STATISTICAL MODELING

NONPARAMETRIC SURE ESTIMATES FOR COMPOUND DECISION PROBLEMS

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

We present a new approach to solving compound decision problems. We first derive the oracle separable decision rule, then directly estimate the rule's unknown parameters by minimizing a SURE estimate of its risk. Unlike existing procedures, we do not assume a parametric or semiparametric prior mixing distribution to derive the form of our estimator. We also do not require nonparametric maximum likelihood estimation of a mixing distribution, which simplifies our theoretical analysis. We apply our estimator to the classical Gaussian sequence problem, show that it can asymptotically achieve the minimum risk among all separable decision rules, and demonstrate its numerical properties in simuations. We next extend our theoretical and numerical results to certain multivariate extensions of the Gaussian sequence problem and apply our method to high-dimensional classification problems in genomics.

✉ sdzhao@illinois.edu

SMOOTHNESS AND SPARSITY ADAPTIVE BAYESIAN TREE ENSEMBLE METHOD FOR HIGH-DIMENSIONAL NONPARAMETRIC REGRESSION

Yun Yang*, *Florida State University*

Antonio Linero, *Florida State University*

Ensembles of decision trees are a useful tool for obtaining for obtaining flexible estimates of regression functions. Examples of these methods include gradient boosted decision trees, random forests, and Bayesian CART. Two potential shortcomings of tree ensembles are their lack of smoothness and vulnerability from curse of dimensionality. We show that these issues can be overcome by instead considering sparsity inducing soft decision trees in which the decisions are treated as probabilistic. We implement this in the context of the Bayesian additive regression trees framework, and illustrate its promising performance through testing on benchmark datasets. We provide strong theoretical support for our methodology by showing that the posterior distribution concentrates at the minimax rate(up-to a logarithmic factor) for sparse functions and functions with additive structures in the high-dimensional regime where the dimensionality of the covariate space is allowed to grow near exponentially in the sample size. Our method also adapts to the unknown smoothness and sparsity levels, and can be implemented by making minimal modifications to existing BART algorithms.

✉ yyang@stat.fsu.edu

SPATIAL ANALYSIS USING SPARSE CHOLESKY FACTORS

Abhirup Datta*, *Johns Hopkins University*

Sudipto Banerjee, *University of California, Los Angeles*

James S. Hodges, *University of Minnesota*

Hierarchical models for regionally aggregated disease incidence data commonly involve region specific latent random effects which are modelled jointly as having a multivariate Gaussian distribution. The covariance or precision matrix incorporates the spatial dependence between the regions. Common choices for the precision matrix include the widely used intrinsic conditional autoregressive model which is singular, and its nonsingular extension which lacks interpretability. We propose a new parametric model for the precision matrix based on a directed acyclic graph representation of the spatial dependence. Theoretical and empirical results demonstrate the interpretability of parameters in our model. Our precision matrix is sparse and the model is highly scalable for large datasets. We also derive a novel order-free version which remedies the dependence of directed acyclic graphs

on the ordering of the regions by averaging over all possible orderings. The resulting precision matrix is still sparse and available in closed form. We demonstrate the superior performance of our models over competing models using simulation experiments and a public health application.

✉ abhidatta@jhu.edu

BAYESIAN MODELING OF INFANT'S GROWTH DYNAMICS WITH PRENATAL EXPOSURE TO ENVIRONMENTAL TOXICANTS

Peter X.K. Song*, *University of Michigan*

Jonggyu Baek, *University of Michigan*

Bin Zhu, *National Cancer Institute, National Institutes of Health*

The early infancy of at-birth to 3 years is critical for infant's developmental tempo and outcomes, which are potentially impacted by prenatal exposure to endocrine disrupting compounds (EDCs). We investigate effects of 10 ubiquitous EDCs on the infant growth dynamics of body mass index (BMI) in a birth cohort study from Mexico City. Modeling growth acceleration is proposed to understand the "force of growth" through a class of semi-parametric stochastic velocity models. The great flexibility of such modeling enables to capture subject-specific dynamics of growth and to assess effects of the EDCs on potential growth delay. We adopted a Bayesian method with the Ornstein-Uhlenbeck process as the prior for the growth rate function, in which the WHO global infant's growth curves were integrated into our modeling and analysis. We found that EDCs exposed during the first trimester of pregnancy were inversely associated with BMI growth acceleration, resulting in a delayed achievement of BMI infancy peak. Such early growth deficiency has been reported as being a profound impact on health outcomes in puberty and adulthood (e.g., timing of sexual maturation).

✉ pxsong@umich.edu

I20. ADVANCES AND INNOVATIVE APPLICATIONS OF JOINT MODELING TO PUBLIC HEALTH RESEARCH

STATISTICAL CHALLENGES IN OBSTETRICS: PREDICTING POOR PREGNANCY OUTCOMES FROM MULTIVARIATE LONGITUDINAL FETAL GROWTH DATA

Paul S. Albert*, *National Cancer Institute, National Institutes of Health*

There are many analytic issues in the area of obstetrics that present challenges for statisticians working in this exciting area. We will focus on the problem of predicting poor pregnancy outcomes such as small for gestational age or preterm birth from longitudinal biomarkers collected during pregnancy. We will begin by presenting a simple two-stage approach that approximates a full maximum-likelihood approach for predicting a biary event from multivariate longitudinal growth data (Albert, Statistics in Medicine, 2012). We will then presnt a more general class of models that accommidates skewed error distributions for the longitudinal data and an asymetric link function for relating the longitudinal data to the binary disease outcome (Kim and Albert, Biometrics, 2016). We will briefly present a tree-based approach for identifying subgroups of women who have an enhanced predictive accuracy for a binary event from fetal growth data (Foster et al., JRSS-A, 2016). Lastly, we will discuss extensions to the case where the outcome is a time to event (gestational age) rather than a binary outcome (Elmi, et al. 2018).

✉ albertp@mail.nih.gov

JOINT ANALYSIS OF MULTIPLE LONGITUDINAL AND SURVIVAL DATA MEASURED ON NESTED TIME-SCALES: AN APPLICATION TO PREDICTING INFERTILITY

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

We consider the joint modeling and prediction of multiple longitudinal processes, a binary process and a skewed continuous longitudinal process and a discrete time-to-event outcome. Data from prospective pregnancy studies provide day-level information regarding the behavior of couples attempting to conceive. It is of considerable interest to develop a risk model for individualized predictions of time-to-pregnancy (TTP) based on behavior, ie, intercourse and menstrual cycle length, a proxy for woman's reproductive health. The intercourse observations are a long series of binary data with a periodic probability of success and the amount of available intercourse data is a function of both the menstrual cycle length and TTP. Moreover, these variables are dependent and observed on different, and nested, time scales (TTP measured in cycles, length of each menstrual cycle on cycles while intercourse measured on days within a menstrual cycle) further complicating its analysis. Here, we propose a semi-parametric shared parameter model for the joint modeling of these processes. Finally, we develop couple-based dynamic predictions to assess the risk for infertility.

✉ sundaramr2@mail.nih.gov

ENAR 2018

PERSONALIZED BIOPSY SCHEDULE FOR PROSTATE CANCER PATIENTS USING JOINT MODELS

Dimitris Rizopoulos*, *Erasmus MC*

Anirudh Tomer, *Erasmus MC*

Ewout Steyerberg, *Erasmus MC*

Monique Roobol, *Erasmus MC*

To reduce the risk of adverse effects from over-treatment, prostate cancer patients are often enrolled in active surveillance (AS) programs. During AS patients are closely monitored to decide when surgery or radiation is to be initiated. However, the success of these programs is greatly compromised by the fact that progression is assessed with prostate gland biopsies. This is an invasive and painful procedure leading to patient non-adherence, reducing therefore the effectiveness of AS. In this work we aim to tackle this problem by replacing the standard biopsies scheme that is the same for all patients, with a custom-made personalized biopsy scheduling procedure that dynamically adapts to the progression rate of a specific patient. More specifically, using data from the PRIAS program, we build joint models to describe the relationship between the progression rate and routinely recorded measurements of serum prostate-specific antigen (PSA) levels. Following, we derive the posterior predictive distribution of progression time, and from it derive the optimal timing for planning the next biopsy.

✉ d.rizopoulos@erasmusmc.nl

DEALING WITH COVARIATES MEASURED AT DEPENDENT FOLLOW-UP VISITS IN THE ESTIMATION OF COX MODELS

Yifei Sun*, *Columbia University*

Chiung-Yu Huang, *University of California, San Francisco*

An important feature of the Cox model is its ability to naturally incorporate time-dependent covariates. In practice, covariate values are measured at intermittent follow-up visits rather than being observed continuously; moreover,

the timing of future follow-up visits may depend on the values of covariates measured at the previous visits as well as the time from the last visits. Simple methods such as carrying forward the last observed covariates in the evaluation of the partial likelihood can result in biased estimation. In this research, we propose a novel semiparametric estimator for evaluating the effects of time-dependent covariates. Instead of postulating a joint model of the covariate processes and the failure time, our approach is based on an estimated score function. Specifically, we kernel smooth the mean functions of weighed covariate processes to derive an asymptotically unbiased estimating equation, where the weight functions depend on the covariate observation time process. The proposed methods are applied to a real data example for illustration.

✉ ys3072@cumc.columbia.edu

I2I. STATISTICS AND INFORMATICS: STRONGER TOGETHER

FOSTERING COLLABORATION BETWEEN BIOSTATISTICS, BIOINFORMATICS, AND BIOMEDICAL INFORMATICS

Lance A. Waller*, *Emory University*

In today's rapidly expanding biomedical data environment, the fields of biostatistics, bioinformatics, and biomedical informatics can often be confused: viewed as synonyms by some, competitors by others. Each offers unique perspectives with increasing needs to manage, capture, and process data of multiple types, sources, and sizes. These needs open the door to increased collaboration between biostatisticians, bioinformaticians, and informaticians. The combined expertise has the power to lead to stronger collaborative research teams that can enhance discovery and translation of biomedical research, but can also lead to counterproductive turf protection. As a Chair of a Department of Biostatistics and Bioinformatics, who also served a brief stint as an Interim Chair of a fledgling Department of Biomedical Informatics, I will share my experiences leading statisticians and informaticians, separately and together. I will highlight future opportunities and challenges as we look to further foster productive collaboration between biostatistics and informatics within the biomedical data sciences.

✉ lwaller@emory.edu

LIFE ON THE EDGE: INTERFACING BIOSTASTISTICS AND BIOINFORMATICS

Kevin R. Coombes*, *The Ohio State University*

As high-throughput experiments in biology become more common, specialized expertise is necessary to organize, integrate, and interpret the resulting data. On the one hand, the size of the data sets has already motivated statistical advances in areas such as multiple testing that are fundamental for "omics" analyses. Newer technologies, like mass cytometry, and modern datasets, like the cancer genome atlas (TCGA) hat contains 10,000 patient samples, pose a different set of statistical challenges. On the other hand, training for bioinformaticians includes an emphasis on databases, data management, and visualization that is unusual in most statistics training. Researchers on both sides of the divide are struggling to find the best ways to perform feature selection and classification to predict rare events. Solving all of these problems requires statistically sound designs and models along with computer science ideas like literate programming and version control to ensure rigor and reproducibility. Using real-life examples, I will discuss how we can integrate ideas from bioinformatics and biostatistics to facilitate rigorous discovery and validation.

✉ coombes.3@osu.edu

USING DATA TO DEFEAT ANTIBIOTIC RESISTANCE

Erinn M. Hade*, *The Ohio State University*

Yuan Gao, *The Ohio State University*

Protiva Rahman, *The Ohio State University*

Courtney Dewart, *The Ohio State University*

Mark Lustberg, *The Ohio State University*

Kurt Stevenson, *The Ohio State University*

Emily Patterson, *The Ohio State University*

Awa Mbodj, *The Ohio State University*

Arnab Nandi, *The Ohio State University*

Courtney Hebert, *The Ohio State University*

Physicians often must prescribe an antibiotic for a suspected infection prior to confirmatory culture results. They balance the risk associated with prescribing a broad antibiotic with the patient side-effects, drug costs, and future resistance. Our work has focused on developing a novel antibiotic advisor that is specific to patient characteristics, and the experience of the hospital. We have become expert in the nuances of the complex microbiology data that informs these models, and of its limitations. We have brought together a team that includes: a statistician, data scientist, clinical informatician, software engineer and physicians to develop and visualize our novel antibiotic advisor. I will discuss our translation of clinical data, informatics tools and statistical methods to develop a robust clinical antibiotic advisor.

✉ hade.2@osu.edu

MOVING A PARALYZED HAND: A BIOMEDICAL BIG DATA SUCCESS STORY

Nicholas D. Skomrock*, *Battelle Memorial Institute*

Michael A. Schwemmer, *Battelle Memorial Institute*

David A. Friedenberg, *Battelle Memorial Institute*

To date, catastrophic spinal cord injury that results in paralysis almost always means that the patient will not be able to control their limbs ever again. However, a recent successful collaboration of statisticians, engineers, computer scientists, and medical doctors is looking to change that through the development of a neural bypass system. The goal of a neural bypass system is to create a brain computer interface that can send signals from the brain to activate muscle movement by bypassing the damaged spinal cord. The resulting interface was the product of creative engineering, high performance computing, massive signal processing, and statistical learning. In this talk, I will describe the brain computer interface that we created and how it led our patient to regain control of his hand. I will also discuss the importance of team science and how it was imperative to the success of the project.

✉ skomrock@battelle.org

I22. RECENT DEVELOPMENTS IN STATISTICAL ANALYSIS OF BRAIN DATA

STATISTICAL MODELING OF BRAIN CONNECTIVITY USING MULTIMODAL NEUROIMAGING

Ying Guo*, *Emory University*

Yingtian Hu, *Emory University*

The study of human brain based on multimodal imaging has become one of the frontiers of neuroscience research. It provides the opportunity to combine modalities to investigate the brain architecture from both functional and structural perspectives. The majority of existing approaches typically examine these modalities in separate analyses, although multimodal methods are emerging to facilitate joint analyses. We present statistical methods for exploring brain organizations by combining information from functional and structural imaging. We will present statistical modeling approaches for helping address the following questions: what is the relationship between structural and functional connections and whether that relationship depends on

subjects' covariates, how to combine the multimodality imaging to advance understanding of brain networks. . We evaluate the performance of the methods through simulation studies and also illustrate their applications in real-world neuroimaging data examples.

✉ yguo2@emory.edu

CALCIUM IMAGING: STATE-OF-THE-ART AND FUTURE CHALLENGES

Jordan Rodu*, *University of Virginia*

Calcium imaging is a powerful technique for observing neuronal populations of ever-increasing sizes, and it creates a unique opportunity for studying population-level dynamics in the brain. Despite the rapid advancement of the technique, many statistical challenges remain. For instance, there is substantial noise introduced in the recordings; images are obtained at a relatively low sampling rate; and there is a nonlinear relationship between the calcium fluorescence and action potentials. These and other factors make spike-timing inference difficult and susceptible to artifacts from pre-processing. We identify major areas of focus and briefly touch on the current state-of-the-art solutions. We also discuss new statistical challenges for calcium imaging studies.

✉ jsr6q@virginia.edu

HUMAN SEIZURES COUPLE ACROSS SPATIAL SCALES THROUGH TRAVELING WAVE DYNAMICS

Mark Kramer*, *Boston University*

Louis-Emmanuel Martinet, *Massachusetts General Hospital*

Emad Eskandar, *Massachusetts General Hospital*

Wilson Truccolo, *Brown University*

Uri Eden, *Boston University*

Sydney Cash, *Massachusetts General Hospital*

Epilepsy - the propensity toward recurrent, unprovoked seizure - is a devastating disease affecting 65 million people worldwide. Understanding and treating this disease remains a challenge, as seizures manifest through mechanisms and features that span spatial and temporal scales. In this talk, we will examine some aspects of this challenge through the analysis and modeling of human brain voltage activity recorded simultaneously across microscopic and macroscopic spatial scales. We will show some evidence that rapidly propagating waves of activity sweep across the cortex during seizure. We will also describe a corresponding computational model to propose specific mechanisms that support the observed spatiotemporal dynamics.

✉ mak@bu.edu

A LOW-RANK MULTIVARIATE GENERAL LINEAR MODEL FOR MULTI-SUBJECT fMRI DATA AND A NON-CONVEX OPTIMIZATION ALGORITHM FOR BRAIN RESPONSE COMPARISON

Tingting Zhang*, *University of Virginia*

Minh Pham, *University of Virginia*

Marlen Z. Gonzalez, *University of Virginia*

James A. Coan, *University of Virginia*

The focus of this paper is on evaluating brain responses to different stimuli and identifying brain regions with different responses using multi-subject, stimulus-evoked functional magnetic resonance imaging (fMRI) data. We present a new low-rank multivariate general linear model (LRMGLM) for fMRI data. The new model not only is flexible to characterize variation in hemodynamic response functions (HRFs) across different regions and stimulus types, but also enables information "borrowing" across voxels and uses much fewer parameters than typical nonparametric models for HRFs. We estimate the proposed LRMGLM through minimizing a new penalized optimization function. We show that the proposed method can outperform several existing voxel-wise methods by achieving both high sensitivity and specificity. We

apply the proposed method to the fMRI data collected in an emotion study and identify anterior dACC to have different responses to a designed threat and control stimuli.

✉ tz3b@virginia.edu

I23. STATISTICAL MODELING TO ADDRESS HUMAN RIGHTS ISSUES

ESTIMATING THE NUMBER OF FATAL VICTIMS OF THE PERUVIAN INTERNAL ARMED CONFLICT, 1980-2000: NEW ANALYSES AND RESULTS

Daniel Manrique-Vallier*, *Indiana University*

Patrick Ball, *Human Rights Data Analysis Group*

David Sulmont, *Pontificia Universidad Catolica del Peru*

We present a new study estimating the the number of people killed or disappeared in the internal armed conflict in Peru between 1980 and 2000. Improving over previous work by the Peruvian Truth and Reconciliation Commission (2003), we take advantage of new available data, as well as recent methodological developments on Bayesian multiple recapture estimation methods.

✉ dmanriqu@indiana.edu

STATISTICS AND JUSTICE: ISSUES IN FORENSIC FEATURE COMPARISONS

Robin Mejia*, *Carnegie Mellon University*

In 2009, the National Research Council released a report calling into question the foundation of many commonly used forensic techniques and outlining a research agenda. In 2016, the President's Council of Advisors on Science and Technology put out a new report raising similar concerns, and in 2017, the National Commission on Forensic Science nearly passed a set of recommendations on statistical statements in court testimony that would have severely limited what forensic analysts could say in court. As the ongoing criticism suggests, research has not kept up with needs in forensics. Statisticians are uniquely positioned to address the problems that plague forensics, which include a need for improved reference data, feature extraction, and comparison and testing methods. In 2015, NIST awarded a $20 million grant to a four-university consortium to address these problems. This talk will highlight new work on improving forensic science. The stake are high. The Innocence Project, an organization that helps wrongfully convicted people. In 46% of over 300 exonerations, it found that the 'misapplication' of forensics contributed to conviction.

✉ rmejia@andrew.cmu.edu

ACCOUNTING FOR RECORD LINKAGE UNCERTAINTY IN POPULATION SIZE ESTIMATION

Mauricio Sadinle*, *University of Washington*

Merging datafiles containing information on overlapping sets of entities is a challenging task in the absence of unique identifiers, further complicated when some entities are duplicated in the datafiles. A Bayesian approach to this problem allows us to incorporate prior information on the quality of the datafiles, which we find to be especially useful when no training data are available, but most importantly provides us with a proper account of the uncertainty about the coreference between records. Here we focus on how to incorporate this coreference uncertainty into a subsequent stage of population size estimation. This two-stage strategy is well suited for when data are subject to privacy constrains and when the final population size model is subject to exploration.

✉ msadinle@gmail.com

THE CAUSAL IMPACT OF BAIL ON CASE OUTCOMES FOR INDIGENT DEFENDANTS

Kristian Lum*, *Human Rights Data Analysis Group*

It has long been observed that defendants who are subject to pre-trial detention are more likely to be convicted than those who are free while they await trial. However, until recently, much of the literature in this area was only correlative and not causal. Using an instrumental variable that

represents judge severity, we apply near-far matching– a statistical methodology designed to assess causal relationships using observational data– to a dataset of criminal cases that were handled by public defenders in a major US city in 2015. We find a strong causal relationship between bail– an obstacle that prevents many from pre-trial release– and case outcome. Specifically, we find setting bail results in a 34% increase in the likelihood of conviction for the cases in our analysis. To our knowledge, this marks the first time matching methodology from the observational studies tradition has been applied to understand the relationship between money bail and the likelihood of conviction.

✉ kl@hrdag.org

I24. LATENT VARIABLES

MIXED MEMBERSHIP REGRESSION MODELS FOR ESTIMATING AUTOIMMUNE DISEASE PATIENT SUBSETS

Zhenke Wu*, *University of Michigan*

Livia Casciola-Rosen, *Johns Hopkins University School of Medicine*

Antony Rosen, *Johns Hopkins University School of Medicine*

Scott L. Zeger, *Johns Hopkins University*

An ongoing challenge in subsetting autoimmune disease patients is how to summarize autoantibody mixture in terms of reactions to elemental sets of autoantigens. In this paper, we develop a model to support this type of substantive research. Our approach is to posit a mixed membership regression model for discovering autoantibody response profiles among patients, in which the mixing weights are parametrized by observed covariates. In this model, the profiles are unobserved multivariate binary presence or absence status measured with error and the profile prevalence is specified as a parsimonious stick-breaking regression model on an arbitrary number of patient-level covariates, such as cancer type, age and time since scleroderma onset, enabling clinicians to introduce elements of phenotypes related to the autoantibody profiles into the model. We demonstrate the proposed method by analyzing patients' gel electrophoresis autoradiography (GEA) data. Our method quantifies the variation of both the frequency and nature of autoantibody profiles by covariates.

✉ zhenkewu@umich.edu

CONSTRUCTING TARGETED LATENT VARIABLES USING LONGITUDINAL DATA TO DEVELOP A MORE SENSITIVE CLINICAL ENDPOINT FOR PROGRESSIVE MULTIPLE SCLEROSIS

Christopher R. Barbour*, *Montana State University*

Mark Greenwood, *Montana State University*

Bibiana Bielekova, *National Institute of Neurological Disorders and Stroke, National Institutes of Health*

Peter Kosa, *National Institute of Neurological Disorders and Stroke, National Institutes of Health*

Scales are often constructed from multiple outcome measures to create a combined metric that better measures the true trait of interest than any of the original components. These methods typically focus on explaining cross-sectional variation in the responses using a projection into a single dimension(s) to define the combinations of variables. When the interest is in creating a scale that is sensitive to changes over time, developing it using cross-sectional data may not tune the projection to detect changes over time optimally. This research develops methodology for scale creation that is optimized to detect variation over time in longitudinal data. An overview of statistical methods traditionally used in scale development will be given. The proposed method, Constructed Composite Response (CCR), will then be presented, and a simulation study performed to examine properties of each projection method across a variety of datasets will be discussed. The method applied to a motivating dataset of multiple sclerosis (MS) patients will highlight strengths and potential improvements to the CCR. Future extensions will be discussed, such as inclusion of sparse-learning strategies.

✉ christopher.barbour@montana.edu

ACKNOWLEDGING THE DILUTION EFFECT IN GROUP TESTING REGRESSION: A NEW APPROACH

Stefani Mokalled*, *Clemson University*

Christopher McMahan, *Clemson University*

Joshua Tebbs, *University of South Carolina*

Christopher Bilder, *University of Nebraska-Lincoln*

From screening for infectious diseases to detecting bioterrorism, group (pooled) testing of bio-specimen is a cost efficient alternative to individual level testing. Group testing has been utilized for both classification (identifying positive individuals) and estimation (fitting regression models using covariate measurements). A concern with the estimation process is the possible dilution of one individual's positive signal past an assay's threshold of detection. To account and correct for this dilution effect, we develop a new group testing regression model which explicitly acknowledges the effect. Unlike previous work in this area, this is accomplished by considering the continuous outcome that the assay measures, the individuals' latent biological marker (biomarker) levels, and the distributions of the biomarker levels of the cases and controls without requiring a priori knowledge of these distributions. We develop a novel mixture model and an expectation-maximization algorithm to complete model fitting. The performance of the methodology is evaluated through numerical studies and is illustrated using Hepatitis B data on Irish prisoners.

✉ smokall@g.clemson.edu

MODELING RATER DIAGNOSTIC SKILLS IN BINARY CLASSIFICATION PROCESSES

Don Edwards*, *University of South Carolina*

Xiaoyan Lin, *University of South Carolina*

Hua Chen, *Oklahoma Medical Foundation*

Kerrie Nelson, *Boston University*

Many disease diagnoses involve subjective judgments by qualified raters. For example, through the inspection of a mammogram, MRI, or ultrasound image, the clinician himself becomes part of the measuring instrument. This paper focuses on a subjective binary classification process, proposing a hierarchical model linking data on rater opinions with patient true disease-development outcomes. The model allows for the quantification of the effects of rater diagnostic skills (bias and magnifier) and patient latent disease severity. A Bayesian Markov chain Monte Carlo (MCMC) algorithm is developed to estimate these parameters. The rater-specific sensitivity and specificity can be estimated using MCMC samples. Cost theory is utilized to identify poor- and strong-performing raters and to guide adjustment of rater bias and diagnostic magnifier to improve the rating performance. The diagnostic magnifier is shown as a key parameter to present a rater's diagnostic ability because a rater with a larger diagnostic magnifier has a uniformly better receiver operating characteristic (ROC) curve when varying the value of diagnostic bias. A simulation study and a mammography example are discussed.

✉ edwards@stat.sc.edu

A GAMMA-FRAILTY PROPORTIONAL HAZARDS MODEL FOR BIVARIATE INTERVAL-CENSORED DATA

Prabhashi W. Withana Gamage*, *Clemson University*

Christopher S. McMahan, *Clemson University*

Lianming Wang, *University of South Carolina*

Wanzhu Tu, *Indiana University School of Medicine*

The Gamma-frailty proportional hazards (PH) model is commonly used to analyze correlated survival data. Despite their popularity, Gamma-frailty PH models for correlated interval-censored data have not received as much attention as other models for right-censored data. In this work, a Gamma-frailty PH model for bivariate interval-censored data is presented and an expectation-maximization (EM) algorithm for model fitting is developed. The proposed model adopts a monotone spline representation for the purposes

of approximating the conditional cumulative baseline hazard functions, significantly reducing the number of unknown parameters while retaining modeling flexibility. The EM algorithm was derived from a novel data augmentation procedure involving latent Poisson random variables. The algorithm is easy to implement, robust to initialization, and enjoys quick convergence. Simulation results suggest that the proposed method provides reliable estimation and valid inference, and is robust to the misspecification of the frailty distribution. To further illustrate its use, the proposed method is used to analyze data from an epidemiological study of sexually transmitted infections.

✉ pwickra@clemson.edu

I25. CAUSAL INFERENCE IN SURVIVAL ANALYSIS

TARGETED MINIMUM LOSS-BASED ESTIMATION WITH INTERVAL-CENSORED TIME-TO-EVENT OUTCOMES

Oleg Sofrygin*, *University of California, Berkeley and Kaiser Permanente*

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

Romain Neugebauer, *Kaiser Permanente*

While formal causal inference methods addressing selection bias from informative right-censoring are widely accepted, similar methods to address bias from informative interval-censoring have received less attention. We revisit this estimation problem by leveraging large-scale EHR data to compare the effectiveness of four personalized time-varying treatment interventions for delaying progression of albuminuria in patients with diabetes. This outcome is defined as the time when the patient's urine albumin-to-creatinine ratio (UACR) crosses a certain level. Because UACR measurements are collected sporadically in clinical settings, the exact timing of this outcome is unknown for most patients. In addition to right-censoring, our outcome of interest is therefore subject to interval-censoring -- we only known if the event occurred between two consecutive UACR measurements. Using both simulated and real data, we demonstrate how traditional estimation approaches that ignore interval-censoring can result in biased estimates. Our simulation study illustrates the potential magnitude of this bias, as well as demonstrates the validity of longitudinal TMLE for interval-censored outcomes.

✉ olegso@gmail.com

INSTRUMENTAL VARIABLE ANALYSIS WITH CENSORED DATA IN THE PRESENCE OF MANY WEAK INSTRUMENTS

Ashkan Ertefaie*, *University of Rochester*

Anh Nguyen, *University of Michigan*

David Harding, *University of California, Berkeley*

Jeffrey Morenoff, *University of Michigan*

Wei Yang, *University of Pennsylvania*

This article discusses an instrumental variable (IV) approach for analyzing censored data that includes many instruments that are weakly associated with the endogenous variable. We study the effect of imprisonment on time to employment using an administrative data on all individuals sentenced for felony in Michigan. Despite the large body of research on the effect of prison on employment, this is still a controversial topic, especially since some of the studies could have been affected by unmeasured confounding. We take advantage of a natural experiment based on the random assignment of judges to felony cases and construct a vector of IVs based on judges' ID that can avoid the confounding bias. However, some of the constructed IVs are weakly associated with the sentence type, which can potentially lead to misleading results. Using a dimension reduction technique, we propose a novel semi-parametric estimation procedure in a survival context that is robust to the presence of many weak instruments. The optimal choice of the test statistic has also been derived. Analyses show a significant negative impact.

✉ ashkan_ertefaie@urmc.rochester.edu

INSTRUMENTAL VARIABLE STRUCTURAL NESTED CUMULATIVE FAILURE TIME MODELS FOR COMPARING THE RISK OF FRACTURE WITH ANTIEPLEPTIC DRUGS

Alisa J. Stephens-Shields*, *University of Pennsylvania*

Xu Han, *Temple University*

Marshall Joffe, *University of Pennsylvania*

Dylan Small, *University of Pennsylvania*

Wei Yang, *University of Pennsylvania*

Instrumental variable (IV) methods are often used to estimate the effects of exposures that may be subject to unmeasured confounding. IV estimators for time-to-event data, which present the analytic challenge of censored outcomes, are not as developed. We propose a flexible Instrumental Variable Structural Nested Cumulative Failure Time Model (IV-SNCFTM) for the causal risk ratio of experiencing a time-to event outcome among alternate levels of an exposure. We estimate model parameters using a semiparametric G-estimation procedure. Our estimator can accommodate time-varying exposures and detect non-constant effects over time. We illustrate our approach using two clinical studies: an observational study of the impact of CYP3A4-inducing vs. CYP3A4 non-inducing antiepileptic drugs (AEDs) on the risk of fracture in The Health Improvement Network, and the Health Insurance Program study, a randomized trial of the effect of breast cancer screening on mortality. We evaluate our estimator through a simulation study considering the impact of instrument strength, censoring, and failure rates, and sample size on performance.

✉ alisaste@pennmedicine.upenn.edu

WEIGHTED ESTIMATORS OF THE COMPLIER AVERAGE CAUSAL EFFECT ON RESTRICTED MEAN SURVIVAL TIME WITH OBSERVED INSTRUMENT-OUTCOME CONFOUNDERS

Sai Hurrish Dharmarajan*, *University of Michigan*

Douglas E. Schaubel, *University of Michigan*

A major concern in observational studies is unmeasured confounding of the relationship between treatment and outcome. Instrumental variable (IV) methods are able to control for unmeasured confounding. However, IV methods developed for censored time-to-event data rely on assumptions that may not be reasonable in practical applications, making them unsuitable for observational studies. In this report, we develop weighted estimators of the complier average causal effect on restricted mean survival time. Our method is able to accommodate instrument-outcome confounding and adjust for covariate dependent censoring, making it particularly suited for causal inference from observational studies. We establish asymptotic properties and derive easily implementable asymptotic variance estimators for the proposed estimators. Through simulations, we show that the proposed estimators tend to be more efficient than propensity score matching based estimators or inverse probability of treatment weighted estimators in certain situations, and tend to perform as well in other situations. We apply our method to compare haemodialysis and peritoneal dialysis modalities using data from the USRDS.

✉ shdharma@umich.edu

WEIGHTED LOG-RANK TESTS ADJUSTED FOR NON-RANDOM TREATMENT ALLOCATION AND DEPENDENT CENSORING

Chenxi Li*, *Michigan State University*

When observational data are used to compare treatment-specific survivals, regular two-sample tests, such as the log-rank test, need to be adjusted for the imbalance between treatments with respect to baseline covariate distributions. Besides, the standard assumption that survival time and censoring time are conditionally independent given the treatment, required for the regular two-sample tests, may not be realistic in observational studies. Moreover, treatment-specific hazards are often non-proportional, resulting in small power for the log-rank test. In this paper, we propose a set of adjusted weighted log-rank tests and their supremum versions by inverse probability of treatment and

censoring weighting to compare treatment-specific survivals based on data from observational studies. These tests are proven to be asymptotically correct. Simulation studies show that with realistic sample sizes and censoring rates, the proposed tests have the desired Type I error probabilities and are more powerful than the adjusted log-rank test when the treatment-specific hazards differ in non-proportional ways. A real data example illustrates the practical utility of the new methods.

✉ cli@epi.msu.edu

MATCHING METHODS FOR EVALUATING THE EFFECT OF A TIME-DEPENDENT TREATMENT ON THE SURVIVAL FUNCTION

Danting Zhu*, *University of Michigan*

Douglas Schaubel, *University of Michigan*

In observational studies of survival time featuring a binary time-dependent treatment, investigators are often more interested in the difference between survival functions instead of hazard ratio. We propose flexible methods applicable to big data sets for the purpose of to estimating the causal effect of treatment among the treated with respect to survival probability. The objective is to compare post-treatment survival with the survival function that would have been observed in the absence of treatment. The proposed methods utilize prognostic scores, but are otherwise nonparametric. Essentially, each treated patient is matched to a group of similar qualified not-yet- treated patients. The treatment effect is then estimated through a difference in weighted Nelson-Aalen survival curves, which can be subsequently integrated to obtain the corresponding difference in restricted mean survival time. Large-sample properties are derived, with finite-sample properties evaluated through simulation. The proposed methods are then, applied to kidney transplantation using data from a national organ failure registry.

✉ dantzhu@umich.edu

I26. ELECTRONIC HEALTH RECORDS

SCALABLE BAYESIAN NONPARAMETRIC CLUSTERING AND CLASSIFICATION

Yang Ni*, *University of Texas, Austin*

Peter Mueller, *University of Texas, Austin*

Yuan Ji, *University of Chicago*

We develop a scalable multi-step Monte Carlo algorithm for inference under (possibly non-conjugate) Bayesian nonparametric models. Each step is "embarrassingly parallel" and can be implemented using the same Markov chain Monte Carlo sampler. The simplicity and generality of our approach makes a wide range of Bayesian nonparametric methods applicable to large datasets. Specifically, we apply product partition model with regression on covariates using novel implementation to classify and cluster patients in large electronic health records study. We find interesting clusters and superior classification performance against competing classifiers.

✉ yangni87@yahoo.com

OUTCOME IDENTIFICATION IN ELECTRONIC HEALTH RECORDS USING PREDICTIONS FROM AN ENRICHED DIRICHLET PROCESS MIXTURE

Bret Zeldow*, *University of Pennsylvania*

Alisa Stephens-Shields, *University of Pennsylvania*

Jason A. Roy, *University of Pennsylvania*

We propose a novel semiparametric model for the joint distribution of a continuous longitudinal outcome and the baseline covariates using an enriched Dirichlet process (EDP) prior. This joint model decomposes into a linear mixed model for the outcome given the covariates and marginals for the covariates. The EDP prior is placed on the regression and spline coefficients, the error variance, and the parameters governing the predictor space. We predict

the outcome at unobserved time points for subjects with data at other time points as well as for completely new subjects with covariates only. We find improved prediction over mixed models with Dirichlet process (DP) priors when there are a large number of covariates. Our method is demonstrated with electronic health records consisting of initiators of second generation antipsychotic medications, which are known to increase the risk of diabetes. We use our model to predict laboratory values indicative of diabetes for each individual and assess incidence of suspected diabetes from the predicted dataset. Our model also serves as a functional clustering algorithm in which subjects are clustered by outcome trajectories.

✉ zeldow@protonmail.com

CHALLENGES IN JOINTLY MODELLING IRREGULAR VISIT PROCESS AND OUTCOME PROCESS IN OBSERVATIONAL STUDIES: REVIEW AND EXTENSION

Janie Coulombe*, *McGill University*

Robert Platt, *McGill University*

Erica E.M. Moodie, *McGill University*

In observational studies using electronic health records (EHRs), the irregularities in frequency and time to next physician visit across and within individuals, and the dependence between a patient's condition and the visit and outcome processes pose challenges in the modelling of the relationship between variables, treatment and outcome. In addition to informative timing of observations, a second challenge that arises with the analysis of EHRs data is the presence of interval censoring, as patient characteristics are measured only at the time they see a doctor. Failing to account for the dependence structure or the censoring in this data may lead to biased estimates of treatment effect (Lipsitz & al., 2002; Finkelstein & al., 2002; Biometrics). In this talk, I will review different methods that have been proposed for modelling irregular outcome-dependent follow-up and propose an extension to the method of Buzkova and Lumley (2009, Stat Med). Our approach will incorporate inverse probability of treatment weights so as to handle unbalanced confounders among the treatment groups, in cases where patients were not randomized to treatment.

✉ janie.coulombe@mail.mcgill.ca

RISK PREDICTION IN CURRENT STATUS DATA USING PARTIAL AND IMPERFECT INFORMATION FROM ELECTRONIC MEDICAL RECORDS

Stephanie Fulane Chan*, *Harvard University*

Xuan Wang, *Harvard University*

Tianxi Cai, *Harvard University*

Electronic medical records (EMRs) are a valuable resource for discovery research. A major problem with EMRs is that disease status is difficult to determine without manual chart reviews, which are very time intensive. Performing these chart reviews gives rise to current status data, a type of survival data where disease status is only determined at one examination time. From the EMRs, we also have access to a wealth of data in addition to the traditional current status data setup; eg. the first occurrence of an ICD9 code for the disease can give us an estimate for survival time. In this paper, we first propose a robust estimator for current status data under a nonparametric transformation model that does not depend on the censoring distribution. We then propose an estimator that incorporates the mismeasured estimates of survival time from the EMRs, obtained by using the derivative of a rank estimator and combining it with our first estimator, and demonstrate that this estimator is more efficient. We illustrate our methods by assessing the effects of genetic markers on coronary artery disease in a cohort of rheumatoid arthritis patients from the Partners HealthCare EMR.

✉ stephaniechan@fas.harvard.edu

ACCOUNTING FOR DEPENDENT ERRORS IN PREDICTORS AND TIME-TO-EVENT OUTCOMES USING VALIDATION SAMPLES AND ELECTRONIC HEALTH RECORDS

Mark J. Giganti*, *Vanderbilt University*

Pamela A. Shaw, *University of Pennsylvania*

Guanhua Chen, *University of Wisconsin, Madison*

Sally S. Bebawy, *Vanderbilt University School of Medicine*

Megan M. Turner, *Vanderbilt University School of Medicine*

Timothy R. Sterling, *Vanderbilt University School of Medicine*

Bryan E. Shepherd, *Vanderbilt University*

Data from electronic health records are prone to errors, which are often correlated across multiple variables. Such errors can have a substantial impact on estimates, yet we are unaware of methods that simultaneously account for errors in covariates and time-to-event outcomes. Using University Care Clinic data, the hazard ratio associated with a 100 cell/mm3 increase in CD4 count was 0.90 (95%CI: 0.82-0.97) using unaudited records and 0.62 (95%CI: 0.54-0.70) using audited records. Our goal is to obtain unbiased and efficient estimates after validating a random subset of records. Treating unvalidated variables as missing, we propose discrete time models built with a validated subsample together with multiple imputation for estimation. Using the complete 100% validated dataset as a gold standard, we compare the mean square error of estimates from our approach with those from the unvalidated dataset and the corresponding subsample-only dataset for various subsample sizes. By incorporating reasonably sized validated subsample data, our approach had improved estimation compared to only using unvalidated data and, under certain conditions, only using the validated subset.

✉ mark.giganti@vanderbilt.edu

I27. METHODS FOR RNA-SEQ DATA

A METHOD FOR MITIGATING THE ADVERSE IMPACT OF BATCH EFFECTS IN SAMPLE PATTERN DETECTION

Teng Fei*, *Emory University*

Tengjiao Zhang, *Tongji University*

Weiyang Shi, *Tongji University*

Tianwei Yu, *Emory University*

Motivation: It is well known that batch effects exist in RNA-seq data and other profiling data. Although some methods do a good job adjusting for batch effects by modifying the data matrices, it is still difficult to remove the batch effects entirely. The remaining batch effect can cause artifacts in the detection of patterns in the data. Results: In this study, we consider the batch effect issue in the pattern detection among the samples. Instead of adjusting the original data matrices, we design an adaptive method to directly adjust the similarity matrix between samples. In simulation studies, the method achieves better results recovering true underlying clusters, compared to the leading batch effect adjustment method ComBat. In real data analysis comparing human and mouse datasets, the method correctly re-aligned samples based on their tissue origin. Availability: The R package is available at: https://github.com/tengfei-emory/QuantNorm.

✉ tfei@emory.edu

CELL TYPE-AWARE DIFFERENTIAL EXPRESSION ANALYSIS FOR RNA-SEQ DATA

Chong Jin*, *University of North Carolina, Chapel Hill*

Wei Sun, *Fred Hutchinson Cancer Research Center*

Mengjie Chen, *University of Chicago*

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

Differential expression using RNA sequencing of bulk tissue samples (bulk RNA-seq) is a very popular and effective approach to study many biomedical problems. However,

most tissue samples are composed of different cell types, presenting challenges to the analysis of bulk RNA-seq, which aggregates gene expression across multiple types of cells. Methods without accounting for cell type composition may mask or even misrepresent important cell type-specific signals, especially for relatively rare cell types. In addition, differential expression using bulk RNA-seq cannot distinguish the effects of differential cell type composition or differential expression of individual cell types. We propose a method to address these limitations: cell type-aware differential expression analysis. Our method tests cell type-specific differential expression using bulk RNA-seq data by incorporating the information of cell type composition, which can be estimated separately using an existing method. We demonstrate the performance of our method in both simulations and real data analysis.

✉ chongjin@live.unc.edu

PennSeq2: EFFICIENT QUANTIFICATION OF ISOFORM-SPECIFIC GENE EXPRESSION FROM RNA-SEQ DATA USING WEIGHTED LIKELIHOOD METHOD

Jian Hu*, *University of Pennsylvania Perelman School of Medicine*

Mingyao Li, *University of Pennsylvania Perelman School of Medicine*

Yu Hu, *University of Pennsylvania Perelman School of Medicine*

The emergence of RNA-seq has provided a powerful tool to study gene regulation at isoform level. Correctly estimating isoform-specific gene expression is important for understanding complicated biological mechanisms and mapping disease related genes. We previously developed PennSeq, a statistical method that estimates isoform-specific gene expression. However, besides the accuracy of estimation, efficiency, which measures the degree of estimation uncertainty, is also an important factor to consider. Unfortunately, few existing methods provide efficiency estimate. Built upon our success on PennSeq, we developed PennSeq2, which employs a weighted likelihood framework to improve estimation uncertainty. For each read, PennSeq2 assigns a weight based on the number of compatible isoforms to assure that more informative reads receive more weight in the likelihood function. It then employs an Expectation-Maximization (EM) algorithm to estimate isoform-specific gene expression. Our simulation results indicate that PennSeq2 not only yields equally accurate isoform expression estimation but also significantly improved estimation efficiency compared to PennSeq and Cufflinks.

✉ jianhu@pennmedicine.upenn.edu

GENE SELECTION AND IDENTIFIABILITY ANALYSIS OF RNA DECONVOLUTION MODEL USING PROFILE LIKELIHOOD

Shaolong Cao*, *University of Texas MD Anderson Cancer Center*

Zeya Wang, *University of Texas MD Anderson Cancer Center*

Wenyi Wang, *University of Texas MD Anderson Cancer Center*

Tumor tissues consist of cancer cells, as well as noncancerous cells. Adjusting for the cellular heterogeneity is critical to biomarker identifications. Recently, many deconvolution methods using gene expression data have been proposed. However, due to natural variation of gene expression profiles, the model identifiability of each gene varies, which in turn biases the estimation of model parameters. Here we propose a profile likelihood based method to adaptively select for most identifiable genes, in order to achieve better model fitting quality and maximal discriminatory power of cell proportion estimation. This method serves as a pre-processing step for running deconvolution methods such as DeMixT. We applied this approach to TCGA RNA-seq data

from 15 cancer types, and observed that tumor purity estimates became more consistent (improvement in Pearson correlation is 0.076) with those estimated using copy number variations, as compared to using a routine pre-processing. In conclusion, systematic gene selection is an important preprocessing step for any deconvolution models and we achieved this goal using a novel profile likelihood derived metric.

✉ scao@mdanderson.org

APPLICATION OF T PRIORS TO SEQUENCE COUNT DATA: REMOVING THE NOISE AND PRESERVING LARGE DIFFERENCES

Anqi Zhu*, *University of North Carolina, Chapel Hill*

Michael I. Love, *University of North Carolina, Chapel Hill*

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

In RNA-seq differential expression analysis, investigators aim to detect genes that change expression level across experimental conditions, despite technical and biological variability in the observations. A basic challenge is to accurately estimate the effect size, often in terms of a logarithmic fold change (LFC) in expression levels. When the counts of sequenced reads are small in either or both conditions, the maximum likelihood estimate for the LFC has high variance, leading to large estimates not representative of true differences. One approach is to use filtering threshold rules and pseudocounts to exclude or moderate estimated LFC. Filtering may result in a loss of genes from the analysis that have true differences across conditions, while pseudocounts do not take into account the statistical information for the LFC for each gene. Here, we propose the use of a wide-tailed prior distribution for effect sizes, which avoids the use of filtering or pseudocounts. The new posterior estimator for LFC is efficient to calculate and has lower bias than previously proposed shrinkage estimators, while still reducing variance for those genes with little statistical information.

✉ anqizhu@live.unc.edu

GLMM-Seq: GENE-BASED DETECTION OF ALLELE-SPECIFIC EXPRESSION BY RNA-Seq

Jiaxin Fan*, *University of Pennsylvania Perelman School of Medicine*

Rui Xiao, *University of Pennsylvania Perelman School of Medicine*

Mingyao Li, *University of Pennsylvania Perelman School of Medicine*

Jian Hu, *University of Pennsylvania Perelman School of Medicine*

Allele-specific expression (ASE) can be quantified by the relative expression of two alleles in a diploid individual, and such expression imbalance may explain phenotypic variation and disease pathophysiology. Existing methods for gene-based ASE detection can only analyze one individual at a time, thus wasting shared information across individuals. To overcome this limitation, we develop GLMM-seq, a generalized linear mixed-effects model that can simultaneously model multi-SNP and multi-individual information. The model is able to detect gene-level ASE under one condition and differential ASE between two conditions (e.g., diseased vs. healthy controls). To model multiple individuals simultaneously, we further extend existing individual-based ASE detection methods using a weighted ordered p-value approach. Extensive simulations indicate that our methods perform consistently well under a variety of scenarios. We further apply our methods to real data in the Genetics of Evoked Response to Niacin and Endotoxemia Study, and our results will provide novel candidates for modulation of innate immune responses in humans.

✉ jiaxinf@pennmedicine.upenn.edu

I28. MULTIPLE TESTING

STATISTICAL CONSIDERATIONS OF MCP-MOD IN APPLYING TO MULTI-REGIONAL DOSE-RESPONSE STUDIES INCLUDING JAPANESE POPULATION

Toshifumi Sugitani*, *Astellas Pharma Inc.*

Yusuke Yamaguchi, *Astellas Pharma Inc.*

MCP-Mod allows for testing the presence of a dose-response signal based on multiple non-linear dose-response models as well as for model-based estimation of targeted doses such as minimum effective dose. Recently, this method has attracted much attention from various stakeholders and has been studied a lot in the literature for its potential to improve efficiency of dose finding throughout drug development. However, when planning a multi-regional dose-response study with Japanese subjects, there seems to exist particular issues due to difference in regulatory requirements. For example, Japanese regulatory agency has accepted confirmatory dose-response studies in many cases to claim efficacy of the test drug, so it would be essential to also include the test of efficacy of individual doses while controlling the familywise error rate. We discuss a promising testing strategy for this purpose and how overall sample size changes compared to the case where no hypothesis testing of efficacy of individual doses is postulated. Moreover, we discuss how dose-response similarity between Japanese and non-Japanese population should be assessed in connection with such a testing strategy.

✉ toshifumi.sugitani@astellas.com

CHOICE OF MULTIPLE COMPARISON PROCEDURES FOR STUDY WITH MULTIPLE OBJECTIVES

Bin Dong*, *Janssen Research & Development*

As multiple comparison procedures are commonly used to control family-wise type I error rate in confirmatory clinical trials, selecting an appropriate MCP that balances the clinical importance of hypotheses and chance of success becomes a key decision in the clinical trial design. Conventionally, MCPs are more often evaluated by marginal power of each hypothesis and/or expected number of rejections under assumed alternative hypotheses. However, the marginal power is conditional to a fixed set of assumptions, and may not fully capture the chance of success, which constitutes the statistical significance of a combination of hypotheses. Therefore, such a chosen MCP can be sub-optimal in addressing the actual trial objective. In this research we investigate the probability of multiple dimensions of rejection regions, utilizing the probabilities, we propose a measure to compare different MCPs with respect to the probability of meeting the study objectives, which can be a combination of multiple hypotheses. We will apply the method to a study design example with multiple endpoints and multiple dose groups.

✉ dongbin1207@hotmail.com

QUASI-BAYESIAN MULTIPLE CONFIDENCE INTERVAL PROCEDURE MAINTAINING FWCR

Taeho Kim*, *University of South Carolina*

Edsel A. Pena, *University of South Carolina*

One way to construct multiple confidence intervals (MCI) is to put different coverage probabilities into the individual confidence intervals so that one can minimize the average interval length, maintaining the global coverage probability over a certain level. However, if we use family-wise coverage rate (MCI version of FWER) for the global coverage probability, the given optimization procedure results in a limited amount of the reduction due to the innate conservativeness of FWCR. For 2000 normal location parameters, for example, the procedure provides only 1.5% reduction compared to the Sidak procedure. To handle this limitation, we introduce prior information for the location parameters. Then, there are cases that the true parameters fall into the one side of the CIs with higher probability. For these cases, we classify the tail of the CIs, removing left or right-tails with respect to a given threshold. This quasi-Bayesian MCI procedure provides tighter average length, maintaining the

frequentist error rate, FWCR. For 2000 normal location parameters with normal priors, the results show about 5% reduction compared to the Sidak procedure in an impartial parameter set up.

✉ taeho@email.sc.edu

A UNIFIED FRAMEWORK FOR WEIGHTED PARAMETRIC MULTIPLE TEST PROCEDURES

Dong Xi*, *Novartis*

Ekkehard Glimm, *Novartis*

Willi Maurer, *Novartis*

Frank Bretz, *Novartis*

We describe a general framework for weighted parametric multiple test procedures based on the closure principle. We utilize general weighting strategies that can reflect complex study objectives and include many procedures in the literature as special cases. The proposed weighted parametric tests bridge the gap between rejection rules using either adjusted significance levels or adjusted p-values. This connection is made by allowing intersection hypotheses of the underlying closed test procedure to be tested at level smaller than τ This may be also necessary to take certain study situations into account. For such cases we introduce a subclass of exact τ-level parametric tests which satisfy the consonance property. When the correlation is known only for certain subsets of the test statistics, a new procedure is proposed to fully utilize this knowledge within each subset. We illustrate the proposed weighted parametric tests using a clinical trial example.

✉ dong.xi@novartis.com

AN EFFICIENT FWER CONTROLLING PROCEDURE FOR DATA WITH REDUCED RANK STRUCTURE

Xing Qiu*, *University of Rochester*

Jiatong Sui, *University of Rochester*

Classical FWER (familywise error rate) controlling procedures have a stigma of being under-powered for high-throughput data. This arguably is the main reason why false discovery rate (FDR) controlling procedures are becoming the default choice for large-scale multiple testing problems. In a recent study, my collaborators and I discovered that if we replace the unrealistic assumption that all hypotheses being tested are independent or weakly dependent by a class of reduced rank correlation structures, we can achieve adequate statistical power and control FWER at a reasonable level simultaneously. In this talk, I will give real data examples (RNA-seq and timecourse microbiome data) that have the reduced rank structure, and show that our proposed procedure, rrMTP, can achieve better statistical power as compared with several other procedures in both simulation studies and two real data applications.

✉ xing_qiu@urmc.rochester.edu

BIAS CORRECTION FOR NONPARAMETRIC TESTS

Duchwan Ryu*, *Northern Illinois University*

Yoonsung Jung, *Prairie View A&M University*

Seong Keon Lee, *Sungshin University*

Nonparametric tests have been developed for past decades to examine the hypothesis without assuming the distributions of test statistics. However, nonparametric location tests are known to be biased under specific circumstances and lose testing power. We examine an example of biased nonparametric scale tests against a scale parameter family, as found in the location test. Further, we consider what causes the bias in the nonparametric tests and propose a nonparametric testing method with an improved power. We present some simulations studies that demonstrate the power improved nonparametric tests and show some applications of the proposed testing method.

✉ dryu@niu.edu

I29. CANCER GENOMICS

UNSUPERVISED CLUSTERING AND VARIABLE SELECTION FOR RNA-Seq DATA

David K.T. Lim*, *University of North Carolina, Chapel Hill*

Naim Rashid, *University of North Carolina, Chapel Hill*

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

Clustering is a form of unsupervised learning that aims to uncover latent groups within data based on similarity across a set of features. A common application of this in biomedical research is in deriving novel cancer subtypes from patient gene expression data, given a set of informative genes. However, it is typically unknown a priori what genes may be informative in discriminating between clusters, and what the optimal number of clusters is. Few methods exist for unsupervised clustering of RNA-seq data that can simultaneously adjust for between-sample normalization factors, account for effects of potential confounding variables, and cluster patients while selecting cluster-discriminatory genes. To address this issue, we propose a mixture model EM algorithm with a group truncated lasso penalty. The maximization is done by coordinate-wise descent using the IRLS algorithm, allowing us to include normalization factors and predictors into our modeling framework. The EM framework allows for subtype prediction in new patients via posterior probabilities of cluster membership given the fitted model. Based on simulations and real data, we show the utility of our method.

✉ deelim@live.unc.edu

STRUCTURED VARIABLE SELECTION OF GENE-ENVIRONMENT INTERACTIONS IN CANCER PROGNOSIS

Guotao Chu*, *Kansas State University*

Yinhao Du, *Kansas State University*

Jie Ren, *Kansas State University*

Cen Wu, *Kansas State University*

In high-throughput profiling studies, identification of gene-environment interactions associated with cancer prognosis has been an important objective. Most of the existing gene-environment interaction methods share the limitation that the "main effect, interaction" hierarchical structure cannot be efficiently identified. In this study, we propose a novel structured variable selection method for hierarchical gene-environment interactions. We impose a set of convex constraints to the penalized loss function and develop a generalized gradient descent based algorithm, which honors the hierarchical structure between main effects and gene-environment interactions. Consequently, the interaction effects will be included only if the main effects are important. Extensive simulation studies have been conducted to demonstrate the advantage of our method over the alternatives in terms of both identification accuracy and prediction performance. Analysis of the Cancer Genome Atlas lung cancer data indicates that gene expressions with important implications have been identified by the proposed method.

✉ chuguotao@ksu.edu

MODELING BETWEEN-STUDY HETEROGENEITY FOR REPRODUCIBLE GENE SIGNATURE SELECTION AND CLINICAL PREDICTION

Naim Rashid*, *University of North Carolina, Chapel Hill*

Quefeng Li, *University of North Carolina, Chapel Hill*

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

In the genomic era, the identification of gene signatures associated with disease is of significant interest. Such signatures are often used to predict clinical outcomes in new patients and aid clinical decision-making. However, recent studies have shown that gene signatures are often not reproducible. This occurrence has practical implications in the generalizability and clinical applicability of such signatures. To improve reproducibility, we introduce a novel approach to select gene signatures from multiple data sets to select

genes whose effects are consistently non-zero by accounting for between-study heterogeneity. We build our model upon robust platform-independent quantities, enabling integration over different platforms of genomic data. A high dimensional penalized Generalized Linear Mixed Model (pGLMM) is used to select gene signatures and address data heterogeneity. We provide asymptotic results justifying the performance of our method relative to other methods and demonstrate its advantage through thorough simulation studies. Lastly, we describe an application of our method to subtype pancreatic cancer patients from four studies using different gene platforms.

✉ naim@unc.edu

IMPACT OF NON-HIERARCHICAL MODEL STRUCTURE AND INHERITANCE MODE ON DETECTING SNP-SNP INTERACTIONS

Hui-Yi Lin*, *Louisiana State University Health Sciences Center*

Po-Yu Huang, *Industrial Technology Research Institute, Taiwan*

Jong Park, *Moffitt Cancer Center and Research Institute*

Although SNP-SNP interaction studies are getting more attention during past decade, related statistical methods are under developed. We previously proposed the SNP Interaction Pattern Identifier (SIPI) approach, a powerful method to detect SNP-SNP interactions by testing the 45 interaction patterns. SIPI takes non-hierarchical models and inheritance modes into consideration. SIPI suffers from a large computation burden. We searched for a mini-version of SIPI, which maintains similar statistical power but reduces computational burden. We tested two candidates: Five-Full and AA9 method. The Five-Full approach tests the five full interaction models with different modes. The AA9 approach considers non-hierarchical interaction models with an additive mode. We compared power of both AA9 and Five-Full with SIPI and other common methods. Our simulation results suggest impact of non-hierarchical model structure is larger than inheritance mode in detecting SNP-SNP interactions. AA9 similar to SIPI is more powerful than Five-Full. Findings in a large-scale prostate cancer study show that AA9 is a powerful and computation-efferent tool for detecting SNP interactions.

✉ hlin1@lsuhsc.edu

MONTE CARLO EXPECTATION MAXIMIZATION ALGORITHM FOR THE HETEROGENEOUS DECONVOLUTION OF MIXED TUMOR EXPRESSION

Rongjie Liu*, *Rice University*

Hongtu Zhu, *University of Texas MD Anderson Cancer Center*

Wenyi Wang, *University of Texas MD Anderson Cancer Center*

Tumor samples normally contain non-unique cell types such as immune or fibroblast cells, which lead to estimation error of gene expression signatures associated with cancer diagnosis and main therapy. The existence of mixed tumor samples consequentially affects gene expression profiling commonly performed using RNA-sequencing. We assume the gene expressions data are available from both unmatched pure normal and mixed tumor tissue samples, where they are assumed to follow negative-binomial (NB) distributions. We hereby propose a method Monte Carlo expectation maximization for the deconvolution (MCEM-DeMix): (i) estimating parameters for normal tissues; (ii) estimating the mixture proportion and parameters for tumor tissues using MCEM; (iii) calculating the normal tissue expression and tumor tissue expression. Simulation studies show that MCEM-DeMix performs well in both deriving the mixture proportion of tumor tissue in various gene expression datasets and calculating the tumor tissue expression with various levels of heterogeneity. The proposed MCEM-DeMix method provides an effective deconvolution way of mixed tumor expression.

✉ rongjie.liu.ee@gmail.com

SEMIPARAMETRIC ANALYSIS OF COMPLEX POLYGENIC GENE-ENVIRONMENT INTERACTIONS IN CASE-CONTROL STUDIES

Alexander Asher*, *Texas A&M University*

Odile Stalder, *University of Bern*

Liang Liang, *Harvard University*

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

Yanyuan Ma, *The Pennsylvania State University*

Nilanjan Chatterjee, *Johns Hopkins University*

Many methods have recently been proposed for efficient analysis of case-control studies of gene-environment interactions using a retrospective likelihood framework that exploits the natural assumption of gene-environment independence in the underlying population. However, for polygenic modelling of gene-environment interactions, a topic of increasing scientific interest, applications of retrospective methods have been limited due to a requirement in the literature for parametric modelling of the distribution of the genetic factors. We propose a general, computationally simple, semiparametric method for analysis of case-control studies that exploits the assumption of gene-environment independence without any further parametric modelling assumptions about the marginal distributions of any of the two sets of factors. The method relies on the key observation that an underlying efficient profile likelihood depends on the distribution of genetic factors only through certain expectation terms that can be evaluated empirically. We develop asymptotic inferential theory for the estimator, evaluate its numerical performance via simulations, and apply it to a study of breast cancer.

✉ alexasher@stat.tamu.edu