1. POSTERS: Statistical Genetics and Genomics

1a. Case-Control Studies of Gene-Environment Interactions: When a Case Might not be the Case

Iryna Lobach*, University of California, San Francisco

Analyses of gene-environment interactions might elucidate biologic mechanisms that underlie complex diseases, such as cancer, diabetes, neurological diseases. Many clinical phenotypes have weak clinico-pathologic correlations therefore the set of cases with the same clinical phenotype might have distinct underlying biologic mechanisms. Misclassification of the disease in case-control studies might pose threat to validity of estimation and inference. We propose a novel analysis based on a pseudolikelihood function where retrospectively collected data are analyzed as a random sample. Simulation experiments and application to a case-control study of Alzheimer's disease demonstrate the proposed method.

EMAIL: iryna.lobach@ucsf.edu

1b. Joint Analysis of Multiple Phenotypes in Association Studies based on Cross-Validation Prediction Error

Xinlan Yang*, Michigan Technological University
Shuanglin Zhang, Michigan Technological University
Qiuying Sha, Michigan Technological University

In genome-wide association studies (GWAS), the joint analysis of multiple phenotypes could have increased power over analyzing each phenotype individually. With this motivation, several methods that combine multiple phenotypic traits have been developed, such as O'Brien's method, Trait-based Association Test that uses Extended Simes procedure (TATES), MAONVA and MultiPhen. However, the performance of these methods under a wide range of scenarios is not consistent: one test may be powerful in some situations, but not in the others. Thus, one challenge in multivariate traits analysis is to construct a test that could maintain good performance across different scenarios. In this article, we propose a novel statistical method to test the association between a genetic variant and multiple phenotypic traits based on cross-validation prediction error (PE). Extensive simulations were conducted to evaluate the type I error rates and to compare the power performance of the PE method with various existing methods. We showed that the PE method controls the type I error rates very well and has consistently high power among all the scenarios we considered.

EMAIL: xinlany@mtu.edu

1c. A Novel Test by Testing an Optimally Weighted Combination of Common and Rare Variants with Multiple Traits

Zhenchuan Wang*, Michigan Technological University
Qiuying Sha, Michigan Technological University
Kui Zhang, Michigan Technological University
Shuanglin Zhang, Michigan Technological University

Pleiotropy, the effect of a single genetic variant or a single gene on multiple traits, has been a widespread phenomenon in complex diseases. In this article, we proposed a test by testing an optimally weighted combination of variants with multiple traits (TOWmuT) to test associations between multiple traits and a weighted combination of variants (both rare and common) in a genomic region. TOWmuT is based on the score test under the linear model, in which the weighted combination of variants is treated as the response variable and multiple traits including covariates are treated as independent variables. The statistic of TOWmuT is the maximum of the score test statistic over different weights. TOWmuT is applicable to different types of traits and can include covariates. Using extensive simulation studies, we compared the performance of TOWmuT with some of the existing methods. Our results showed that, in all of the simulation scenarios, TOWmuT is either the most powerful test or comparable to the most powerful test among the tests we compared.

EMAIL: zwang10@mtu.edu

1d. Detection of Differentially Methylated Regions with Generalized Functional Linear Model

Hongyan Xu*, Augusta University
Varghese George, Augusta University

DNA methylation plays important role in cell development and differentiation. Abnormal methylation has been involved

in many diseases especially in cancers. Because of the local correlations of methylation rate across CpG sites in a genomic region, it is of great interest to identify differentially methylated regions (DMRs) between different biological conditions. Many statistical methods have been developed such as Bumphunter and BiSeq. In this study, we propose a method to detect DMRs based on a generalized functional linear model. Simulations show that the test statistic is more powerful than Bumphunter and BiSeq, while having proper control of type I error. Our method accounts for the effects of potential confounders such as age, sex, and cell types naturally by including them as covariates in the model. We illustrate our method in an application to detect DMRs in a study of chronic lymphocytic leukemia.

EMAIL: hxu@augusta.edu

1e. A Method for Joint Processing of Mass Spectrometry-Based Metabolomics Data for Improved Differential Analysis

Leslie Myint*, Johns Hopkins Bloomberg School of Public Health
Kasper Daniel Hansen, Johns Hopkins Bloomberg School of Public Health

As mass spectrometry-based metabolomics becomes a more popular means of scientific investigation, it is important that fundamental data processing and analysis paradigms be revisited to ensure high quality inference. Data preprocessing efforts have largely focused on algorithms that process samples individually and subsequently attempt to group corresponding peaks back together. We show that this technique leads to unnecessary variability in peak quantifications that hurts downstream analysis. We present a new method, bakedpi, that relies on an intensity-weighted bivariate kernel density estimate on a pooling of all samples to detect peaks. When comparing to the widely-used sample-specific processing method XCMS, we see that for all 10 datasets examined, over 50% of overlapping peaks have lower residual standard deviation when quantified with our method than with XCMS. The majority of datasets show lower variability for over 75% of overlapping peaks. We also show that bakedpi has higher power than XCMS. Our results suggest that differential analysis in untargeted comparative metabolomics studies benefit from jointly preprocessing samples before statistical analysis.

EMAIL: leslie.myint@gmail.com

1f. Empirical Estimation of Sequencing Error Rates Using Smoothing Splines WITHDRAWN

Xuan Zhu*, University of Texas MD Anderson Cancer Center

Next-generation sequencing has been used to address a diverse range of biological problems through. However, compared to conventional sequencing, the error rates for next-generation sequencing are often higher, which impacts downstream genomic analysis. Recently, Wang et al. (2012) proposed a shadow regression approach to estimate error rates for next-generation sequencing based on assumption of a linear relationship. However, this linear relationship may not be appropriate for all types of sequence data. Therefore, it is necessary to estimate the error rates in a more reliable way without assuming linearity. We proposed an empirical error rate estimation approach that employs cubic and robust smoothing splines to model the nonlinear relationship. We performed simulation studies using a frequency-based approach to mimic real data structure. The proposed approach provided more accurate estimations than shadow regression approach for all simulation scenarios tested. We also applied the proposed approach for MAQC project, a mutation screening study, Encyclopedia of DNA Elements project, and bacteriophage PhiX DNA samples.

EMAIL: xuan3729@gmail.com

1g. Varying Index Coefficient Model for Dynamic Gene-Environment Interactions

Jingyi Zhang*, Michigan State University
Xu Liu, Shanghai University of Finance and Economics
Yuehua Cui, Michigan State University

Gene-environment interactions play key roles in human complex diseases. Existing literature has shown the power of integrative gene-environment interaction analysis by considering the joint effect of environmental mixtures. In this work, we propose a varying index coefficient model for multiple longitu-

dinal measurements of environmental variables and assess how the genetic effects on a disease trait are nonlinearly modified by a mixture of environmental influences. We derive an estimation procedure for the nonparametric varying index coefficients based on the quadratic inference functions and penalized splines. Theoretical results such as consistency and asymptotic normality of the estimates are established. We also evaluate the performance of our estimation procedure through Monte Carlo simulation studies. In addition, we propose a hypothesis testing procedure for the nonparametric index function in order to investigate the linearity of G-E interactions.

EMAIL: zhang317@stt.msu.edu

1h. AC-PCA: Simultaneous Dimension Reduction and Adjustment for Confounding Variation

Zhixiang Lin*, Stanford University
Can Yang, Hong Kong Baptist University
Hongyu Zhao, Yale University
Wing Hung Wong, Stanford University

Dimension reduction methods are commonly applied to high-throughput biological datasets. However, the results can be hindered by confounding factors, either biological or technical in origin. In this study, we extend Principal Component Analysis to propose AC-PCA for simultaneous dimension reduction and Adjustment for Confounding variation. We show that AC-PCA can adjust for a) variations across individual donors present in a human brain exon array dataset, and b) variations of different species in a model organism ENCODE RNA-Seq dataset. Our approach is able to recover the anatomical structure of neocortical regions, and to capture the shared variation among species during embryonic development. For gene selection purposes, we extend AC-PCA with sparsity constraints, and propose and implement an efficient algorithm. The methods developed in this paper can also be applied to more general settings.

EMAIL: zl235@stanford.edu

1i. Modelling Tumor-Specific eQTL in the Presence of Infiltrating Cells

Douglas R. Wilson*, University of North Carolina, Chapel Hill
Wei Sun, Fred Hutchinson Cancer Research Center
Joseph Ibrahim, University of North Carolina, Chapel Hill

The study of expression quantitative trait loci (eQTL) has illuminated the functional roles of genetic variants associated with many diseases. Computational methods have been developed for eQTL mapping using expression data from microarray or RNA-seq. Application of these methods to study the genetic basis of expression from tumor tissues is problematic as gene expression of tumor tissues represents mixtures of signals from tumor and infiltrating normal cells such as immune cells. We developed a new method for eQTL mapping using RNA-seq data accounting for the variation of tumor purity across samples. Tumor purity of a sample is defined as the proportion tumor cells among all the cells from this tumor sample. Our method separately estimates the eQTL effects on gene expression from tumor and infiltrating normal cells. We demonstrate that our method controls type I error and has higher power than some naïve approaches. We applied our method to study RNA-seq data from The Cancer Genome Atlas and identified some interesting cases where eQTL effect sizes are different between tumor cells and infiltrating normal cells.

EMAIL: drwilson@email.unc.edu

1j. Bivariate Analysis of Genetic Effects on AMD Progression with Intermittent Assessment Times

Tao Sun*, University of Pittsburgh
Wei Chen, Children's Hospital of Pittsburgh of UPMC
Ying Ding, University of Pittsburgh

Age-related macular degeneration (AMD) is the major cause of blindness for the elderly population in the developed countries. It is known as a progressive neurodegenerative disease. However, the genetic causes on its progression have not been elucidated. Using data from a large randomized trial, Age-Related Eye Disease Study (AREDS), where the participants have been followed every 6-12 months for up to 12 years, we aim to evaluate the effects of 34 known AMD risk variants on the disease progression. In doing so, we derived the time inter-

vals for progression-to-late-AMD in both eyes based on their assessment times and modeled them using a novel bivariate survival approach, appropriately accounting for between-eye correlation with the interval censored time-to-event data. In addition, we derived a genetic risk score (GRS), a weighted average effect of these 34 risk variants, and analyzed its effect on AMD progression. We found that GRS was significantly associated with AMD progression (Hazard ratio (HR) =1.39, 95% CI: 1.31-1.48, p=5.38E-27) and the most significant variant for AMD progression was from locus ARMS2/HTRA1 (rs3750846) with an estimated HR of 1.55 (p=1.82E-13).

EMAIL: tao.sun@pitt.edu

1k. Inferring Intra-Tumor Heterogeneity by Jointly Modeling Copy Number Aberrations and Somatic Point Mutations

Chong Jin*, University of North Carolina, Chapel Hill
Wei Sun, Fred Hutchinson Cancer Research Center
Mengjie Chen, University of North Carolina, Chapel Hill

A tumor sample within a single patient can be a conglomerate of heterogeneous cells. Understanding intra-tumor heterogeneity may help us to identify useful biomarkers to guide the practice of precision medicine. We have developed a new statistical method that 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. Cellular prevalence, allele-specific copy number and multiplicity of point mutations within each subclone can be estimated by maximizing the model likelihood. We applied our method to infer the clonal composition in tumor tissues from TCGA colon cancer samples.

EMAIL: chongjin@live.unc.edu

1l. Testing Differential Networks with Application to the Detection of Microbial Interactions

Jing Ma*, University of Pennsylvania
Hongzhe Li, University of Pennsylvania
Tony T. Cai, University of Pennsylvania
Yin Xia, University of North Carolina, Chapel Hill

Microorganisms such as bacteria do not exist in isolation but form complex ecological interaction networks. Conventional methods such as Gaussian graphical models cannot be used to study the conditional independence among bacterial taxa, because the measurements taken are of relative bacterial abundances which are not Gaussian distributed. Alternatively, the Ising model can be used for modeling the conditional independence structure for binary data if one is only interested in the dependency structure among presence/absence of certain bacterial taxa. We propose a testing framework based on the Ising model to detect whether the microbial network structures under different treatment conditions are the same. If they are not the same, our method also conducts multiple testing of network interactions to detect microbe-microbe interactions associated with a binary trait such as presence of a disease.

EMAIL: jinma@upenn.edu

1m. Challenges and Solutions for Analyzing Single Cell Methylation Data

Divy S. Kangeyan*, Harvard University
Martin J. Aryee, Harvard University

Recently developed single cell DNA methylation analysis assays have promised to reveal patterns of inter-cellular epigenetic heterogeneity that have implications on both normal and disease development. Since single cell DNA methylation is a nascent technology, data obtained from it has several unique characteristics including binary methylation status and large number of missing methylation status. To address these issues and current lack of suitable analysis tools, we have developed a computational pipeline for single cell methylation assays. In order to address the sparse nature of the single cell methylation data, we reduced the dimensions from CpG-level data into biologically meaningful features via genomic annotations or variably methylated regions (VMRs). We show results of the method applied to Reduced Representation Bisulfite Sequencing (RRBS) data sets from two different settings: early embryonic mouse stem cells and human chronic lymphocytic leukemia (CLL) cells. In both data sets we were able to see variation in the methylation landscape at cellular level and

using biologically relevant dimension reduction led to better identification of cellular identity.

EMAIL: divyswar01@g.harvard.edu

1n. Super-Delta: A New Approach that Combines Microarray Data Normalization and Differential Expression Test

Yuhang Liu*, Florida State University
Xing Qiu, University of Rochester
Jinfeng Zhang, Florida State University

Data normalization is crucial to gene expression analyses by removing systematic noises. A main drawback of variance reduction is that it borrows information from all genes, which includes differentially expressed genes (DEGs). Such practice will inevitably introduce bias, resulting in inflate of type I error and reduction of power. In this study, we propose a new differential expression analysis pipeline, dubbed as super-delta. This procedure involves a robust strategy to exclude genes with large group difference for normalization, followed by a modified t-test based on asymptotic theory. We compared super-delta with three commonly used normalization methods: global, median-IQR, and quantile normalization, by applying all four methods to a microarray dataset on breast cancer patients who took chemotherapy. Super-delta consistently identified more DEGs with biological connections to breast cancer or chemotherapy, verified by functional enrichment analyses. Simulations showed that super-delta had better statistical power with tighter type I error control than its competitors. In many cases, the performance of super-delta was close to an oracle test using noise-free datasets.

EMAIL: fhlsjs@gmail.com

2. POSTERS: High-Dimesional Data and Variable Selection Methods

2a. Three-Dimensional Data Exploration Based on Body-Centered Cube Lattices

Daniel B. Carr, George Mason University
Zijing Zhang*, George Mason University

With large data sets in three-dimensional space, the number of plots to view massive data increases dramatically. Cognostics, a computer guiding diagnostics, provides a way of prioritizing a large number of plots for interesting patterns. With focus in 3-D extension of cognostics, the goal of this work is to contribute graphical methods that address some of the problems and common tasks such as identifying outliers and tail structure, assessing and comparing central structures. The graphical methods handle large data sets problem in 3-D by binning with truncated octahedron that is the nearest neighbor region of the body-centered cube lattice, and adapting image processing algorithms such as erosion and dilation. An R package is subsequently created for the visualization of massive data in 3-D. Applications of methods in different fields are demonstrated. The work in the area of 3-D medical imaging is under exploration.

EMAIL: zzhang13@gmu.edu

2b. A Split-and-Merge Approach for Singular Value Decomposition of Large-Scale Matrices

Faming Liang, University of Florida
Runmin Shi*, University of Florida
Qianxing Mo, Duncan Cancer Center

We have proposed a new SVD algorithm based on the split-and-merge strategy, which possesses an embarrassingly parallel structure and thus can be efficiently implemented on a distributed or multicore machine. The new algorithm can also be implemented in serial for online eigen- analysis. The new algorithm is particularly suitable for big data problems: Its embarrassingly parallel structure renders it usable for feature screening, while this has been beyond the ability of the existing parallel SVD algorithms.

EMAIL: shirunmin@foxmail.com

2c. Expected Conditional HSIC for Testing Independence

Chenlu Ke*, University of Kentucky
Xiangrong Yin, University of Kentucky

We propose a novel conditional version of Hilbert-Schmidt Independence Criterion for testing independence between two random vectors. We study a specific new index by using Gauss-

ian kernel, which also has an interpretation in terms of a weighted distance between characteristic functions. Two empirical estimates and corresponding independence tests are developed under different scenarios. We illustrate the advantages of our methods in comparing to other existing methods by simulations across a variety of settings and real data applications.

EMAIL: chenlu.ke@uky.edu

2d. Comparison of Treatment Regime Estimation Methods Incorporating Variable Selection: Lessons from a Large Simulation Study

Adam Ciarleglio*, Columbia University
and New York State Psychiatric Institute
Eva Petkova, New York University
Thaddeus Tarpey, Wright State University
Todd Ogden, Columbia University

Increased emphasis on precision medicine has prompted the development of myriad statistical methods for both identifying moderators of treatment effect and estimating optimal rules for selecting treatment. Though many of these methods are promising, they do not perform equally well in all settings. Currently there is little guidance as to which method one should choose to use in practice. In this talk, we present results from a large simulation study in which we compare different recently-developed methods that simultaneously select important moderators and estimate treatment decision rules in a wide variety of realistic settings. The methods are compared with respect various performance metrics and guidelines for selecting methods to use in practice are proposed.

EMAIL: ciarleg@nyspi.columbia.edu

2e. Quantile-Based Subgroup Identification in Randomized Clinical Trials

Youngjoo Cho*, University of Rochester Medical Center
Debashis Ghosh, University of Colorado, Denver

In randomized clinical trials, the effects of treatment are quite heterogeneous, and covariates may be used to stratify patients into subgroups. From a decision-making point of view, we argue that it might be of interest to identify subgroups by using quantiles of outcome of interest. We exploit the potential outcomes framework, in conjunction with quantile regression approaches for subgroup identification. Various approaches using the least absolute selection and shrinkage operator in conjunction with stability selection are considered. Simulation studies and a real data example are used to demonstrate the methodology.

EMAIL: youngjoo_cho@urmc.rochester.edu

2f. Identifying Number of Cells to Consider for Downstream Analysis in Drop-Seq Single Cell RNA Sequencing

Julius S. Ngwa*, Johns Hopkins Bloomberg School of Public Health
Melissa Liu, Johns Hopkins University School of Medicine
Robert Wojciechowski, Johns Hopkins University School of Medicine
Terri Beaty, Johns Hopkins Bloomberg School of Public Health
Don Zack, Johns Hopkins University School of Medicine
Ingo Ruczinski, Johns Hopkins Bloomberg School of Public Health

Single-cell RNA-Seq is widely used for transcription profiling of individual cells. Drop-Seq offers exciting possibilities for capturing single-cells. One key question is estimating number of cells for downstream analysis. Macosko provides one approach by considering reads-per-cell and selecting an inflection point from a cumulative distribution of reads. Barcodes to the right of inflection point are considered empty beads possibly exposed to ambient RNA. In instances where inflection point is not evident, estimating the number of cells becomes challenging. We analyzed expression profiles of retinal mouse cells captured using Drop-Seq. We explored the number of reads and difference in reads-per-barcode. We compared FDR p-values and trajectory of top 50 genes differentially expressed with varying cells. Our data showed the number of cell barcodes to the left of inflection point was not clearly evident. As number of cells decreased the p-value became less significant due to decrease in power. In Drop-Seq, ambiguity in transitioning from beads sampled in cellular RNA to ambient RNA beads can result in false differential expression calls and cell type misclassification.

EMAIL: jngwa1@jhu.edu

2g. High Dimensional Mediation Analysis with Latent Factors

Andriy Derkach*, National Cancer Institute, National Institutes of Health
Joshua Sampson, National Cancer Institute, National Institutes of Health
Ting Huei Chen, Laval UniversityRuth Pfeiffer, National Cancer Institute, National Institutes of Health

Modern biomedical and epidemiological studies often measure a large number of biomarkers such as gene expression and metabolite levels. These biomarkers may be mediators, explaining the relationship between an exposure and an outcome. Standard methods in mediation analysis and causal inference are available for evaluating whether a single variable is a mediator. However, methods to simultaneously assess multiple mediators have received limited attention. Here we propose to jointly model a biological pathway between an exposure. latent mediators affecting multiple biomarkers, and an outcome. To ensure that latent factors influence only a sparse set of biomarkers, we incorporate L1-penalties when fitting the joint model. We extend existing E-M algorithms for factor analysis to incorporate penalty functions and the joint relationship between the exposure, factors, and outcome. We show convergence of E-M algorithm and consistency of estimates. Our proposed methodology can accommodate non-Gaussian outcomes, retrospective sampling. We evaluate the performance of this methodology in small samples through extensive simulations.

EMAIL: andriy.derkach@nih.gov

2h. Linked Matrix Factorization

Michael J. O'Connell*, University of Minnesota
Eric F. Lock, University of Minnesota

In modern biomedical studies, it is common to collect multiple high-throughput datasets that are distinct but interrelated. There have been several methods developed in recent years for decompositions of multi-source data, consisting of more than one matrix with at least one shared dimension. In particular, several methods have been developed for the simultaneous dimension reduction and decomposition of multiple matrices. Typically, these methods assume that just one dimension (rows or columns) is shared among sources. In particular, they assume that either features are shared for different sample sets or that samples are shared for different feature sets. However, these algorithms do not allow for simultaneous horizontal and vertical integration. For data sets that have shared sample sets and shared feature sets, we developed the Linked Matrix Factorization algorithm (LMF), an alternating least squares-based method that allows for decomposition of three matrices simultaneously when one matrix shares its sample set with one matrix and its feature set with another. We illustrate the application of LMF with a high-throughput toxicological screening experiment.

EMAIL: oconn725@umn.edu

2i. A Model-Free Approach to Estimate the Skeletons of High Dimensional Directed Acyclic Graphs

Jenny Yang*, University of North Carolina, Chapel Hill
Wei Sun, Fred Hutchinson Cancer Research Center

Advances in -omics (e.g., measurement of genome-wide genetic variants (genomics) or gene expression (transcriptomics) promise precision medicine: medical practice tailored to the individual patient characterized by the -omic data collected from the patient. To fulfill such a promise, it is crucial to develop statistical methods that effectively model the high-dimensional -omic data and allow actionable conclusions for medical practice. Graphical models, such as the directed acyclic graph (DAG), are among the most promising solutions. Traditionally methods for DAG inference rely on the multivariate Gaussian or structural equation models. Non-linear relations cannot be effectively modeled in these existing methods. We propose a model free approach to estimate the skeleton (i.e., the undirected version) of high dimensional DAGs in two steps. First, estimate the moral graph using penalized model-free variable selection method; then, remove false connections in a moral graph using nonparametric testing. We study the asymptotic properties of our method and demonstrate its advantage in both simulations and real data analysis of gene expression data from breast cancer patients.

E-MAIL: jhyang@live.unc.edu

3. POSTERS: Bayesian Methods

3a. Bayesian Hierarchical Group Spectral Clustering for Brain Functional Connectivity

Eunjee Lee*, University of Michigan
Joseph G. Ibrahim, University of North Carolina, Chapel Hill
Hongtu Zhu, University of Texas MD Anderson Cancer Center

Resting state functional connectivity of human brain is a promising biomarker in brain disorders accompanying neurodegeneration. But standard approaches in functional connectivity studies suffer from a high-dimensional and spatial structure of functional connectivity. We propose a Bayesian group hierarchical spectral clustering model to examine if the functional connectivity is disrupted in subjects with brain disorders. Our method decomposes the functional connectivity matrices into an underlying relational structure among brain areas and subject-specific network in a low-dimensional space. An additional layer is added in our model to incorporate effects of other covariates, which enables to test the group difference of functional connectivity. We take a Bayesian approach to estimate parameters in our model. Our real data analysis revealed that bilateral precuneus had weaker connection between right posterior cingulate gyrus and left angular gyrus for Alzheimer's disease patients than mild cognitive impairment (MCI) patients. There was connectivity difference of paracentral gyrus with other brain regions including superior, middle, inferior frontal gyri.

EMAIL: eunjee@umich.edu

3b. Optimal Point Estimates and Credible Intervals for Ranking County Health Indices

Patricia Jewett, University of Wisconsin, Madison
Ronald Gangnon*, University of Wisconsin, Madison

It is fairly common to rank different geographic units, e.g. counties in the United States, based on health indices. In a typical application, point estimates of the health indices are obtained for each county, and the indices are then simply ranked as if they were known constants. Several authors have considered optimal rank estimators under squared error loss on the rank scale. While computationally convenient, squared error loss on the rank scale may not represent the true inferential goals of rank consumers. We construct alternative loss functions based on three components: (1) the inferential goal (rank position or pairwise comparisons), (2) the scale (original, log-transformed or rank) and (3) the (positional or pairwise) loss function (0/1, squared error or absolute error). We can obtain optimal ranks for loss functions based on rank positions and nearly optimal ranks for loss functions based on pairwise comparisons paired with highest posterior density (HPD) credible intervals. We compare inferences produced by the various ranking methods, both optimal and heuristic, using low birth weight data for counties in the Midwestern United States, 2006-2012.

EMAIL: ronald@biostat.wisc.edu

3c. Full Bayesian Estimation Under Informative Sampling

Luis Gonzalo Leon Novelo*, University of Texas Health Science Center at Houston
Terrance Savitsky, U.S. Bureau of Labor Statistics
Emily V. Leary, University of Missouri

Bayesian estimation is increasingly popular for performing model-based inference to support policy-making. These data are often collected from surveys under informative sampling designs where subject inclusion probabilities are designed to be correlated with the response variable of interest. Survey weights constructed from marginal inclusion probabilities are typically used to form a plug-in, pseudo posterior estimator to approximate the population posterior distribution. We propose a fully Bayesian alternative that jointly models the weights and response under the joint distribution for population generation and the taking of a sample. Our approach is very general and allows for unbiased inference under any model for the population specified by the data analyst, while accounting for all sources of uncertainty in the joint distribution that the plug-in estimator does not. We explore required assumptions to guarantee that our model, estimated on quantities solely observed in the sample, is consistent for the population distribution.

EMAIL: luis.g.leonnovelo@uth.tmc.edu

3d. Bayesian Approach to Misclassification with Partially Validated Data

Katrina J. Anderson*, Baylor University
James D. Stamey, Baylor University

Misclassification of epidemiologic and observational data is a problem that commonly arises and can have adverse ramifications on the validity of results if not properly handled. Considerable research has been conducted when only the response or only the exposure are misclassified, while less work has been done on the simultaneous case. We extend previous frequentist work by investigating a Bayesian approach to dependent, differential misclassification models. Using a logit model with misclassified binary response and exposure variables and assuming a validation sub-sample is available, we compare the resulting confidence and credible intervals under the two paradigms. We compare the results under varying percentages of validation subsamples, 100% (ideal scenario), 25%, 15%, 10%, 5%, 2.5%, and 0% (naïve scenario) of the overall sample size.

EMAIL: katrina_anderson@baylor.edu

3e. Direct and Indirect Comparison of Treatments using Mixtures of Finite Polya Tree Estimation of the AUC

Johanna S. Van Zyl*, Baylor University
and Baylor Institute of Immunology Research
Jack D. Tubbs, Baylor University

A Bayesian nonparametric model using mixtures of finite Polya trees (MFPT) is used to evaluate the effectiveness of two treatment arms from separate studies with a comparable control group through direct and indirect comparisons. MFPT modeling allows the flexibility to estimate densities and distributions that do not exhibit unimodal symmetric shapes. The receiver operating curve (ROC) is estimated using the survival curves and numerical integration is used to obtain an estimate of the area under the ROC curve (AUC). Simulations are used to compare the approach based on the MFPT model with a Bayesian binormal modeling approach where the AUC have a closed-form expression. The method is illustrated using a scenario simulated based on summary statistics from clinical trials.

EMAIL: Johanna_Van_Zyl@baylor.edu

3f. Bayesian Regression of Group Testing Data

Christopher S. McMahan*, Clemson University
Joshua M. Tebbs, University of South Carolina
Christopher R. Bilder, University of Nebraska, Lincoln

Group testing involves pooling individual specimens and testing the pools for the presence of a disease. When individual covariate information is available, a common goal is to relate an individual's true disease status to the covariates in a regression model. Estimating this relationship is a nonstandard problem in group testing because not all individual responses are necessarily observed and all testing responses are subject to misclassification arising from assay error. Previous regression methods for group testing data can be inefficient because they are restricted to using only initial pool responses and/or they make potentially unrealistic assumptions regarding the assay error probabilities. To overcome these limitations, we propose a general Bayesian regression framework for modeling group testing data. The novelty of our approach is that it can be easily implemented with data from virtually any group testing algorithm. Furthermore, our approach will simultaneously estimate assay error probabilities and can even be applied in disease screening situations where multiple assays are used. We illustrate our methods using chlamydia and gonorrhea data.

EMAIL: mcmaha2@clemson.edu

3g. Robust Model-Based Clustering from Multivariate and Grouped Data via Local Deviation Processes

Briana J.K. Stephenson*, University of North Carolina, Chapel Hill
Amy H. Herring, University of North Carolina, Chapel Hill
Andrew Olshan, University of North Carolina, Chapel Hill

Model-based clustering is often applied to multivariate data with systematic differences across groups. These models often realize a growing number of clusters that expand with the dimensionality of the dataset, leading to a loss in cluster interpretability and oversensitivity to deviations among groups. Our goal is to develop a robust and parsimonious clustering method to address these complexities. Traditionally, clustering methods assume subjects allocated to the same

cluster will respond identically to all responses, aberrations between some responses can yield valuable information. Motivated from the local partition process framework, we propose a new method called Robust Profile Clustering (RPC) that allows subjects to aggregate at two levels: 1) globally, where subjects allot to a single cluster via standard Dirichlet process and 2) locally, where individual responses deviate from global indicators via a Beta-Bernoulli process to adapt for differences across groups. Using data from the National Birth Defects Prevention Study, we apply this method to derive dietary patterns of pregnant women in the United States, while adjusting for potential state-level differences.

EMAIL: bjks@live.unc.edu

3h. Parameterizing Heavy-Tailed Hierarchical Survival Models for Hamiltonian Monte Carlo

Krzysztof M. Sakrejda*, University of Massachusetts, Amherst
Nicholas G. Reich, University of Massachusetts, Amherst

Flexible massively hierarchical Bayesian survival models can be used to reduce the noise in estimates of survival time for small batches in large populations. These models have been narrowly applied due to computational difficulties. Real-world data often includes non-ignorable outliers making which further complicate estimation. Auto-tuned Hamiltonian Monte Carlo has made massively hierarchical models practical by reducing tuning and computation time; more importantly these models have become more accessible to non-specialists due to standard diagnostic practices and reparameterizations for troubleshooting. Such practical guidance is not available for important survival distributions. We explore the generalized gamma family of models in the context of HMC and deep hierarchical model structure using simulated data as well as infectious disease reporting delay data as a case study. We consider estimates for data simulated from each model fit to both true and mis-specified models. We also assess the effect of parameterization on the frequency of numerical problems and the quality of convergence diagnostics.

EMAIL: krzysztof.sakrejda@gmail.com

3i. Hierarchical Bayesian Method for Dealing with Uncertainty of Dose-Response Relationship Estimation in Meta-Analysis

Deukwoo Kwon, University of Miami and Sylvester Comprehensive Cancer Center
Jeesun Jung, National Institute on Alcohol Abuse and Alcoholism, National Institutes of Health
Isildinha Reis*, University of Miami and Sylvester Comprehensive Cancer Center

In many clinical trials evaluating radiation therapy, limited number of patients per study and small number of toxicity events make the elucidation of dose-response relationship uncertain. When combining data from multiple studies (meta-analysis), bootstrap approach is applied to assess the uncertainty of the estimated summary population-based dose-response relationship. This approach accounts for sample variability but ignores heterogeneity among studies. We propose a hierarchical Bayesian method which takes into account within and between studies variability. We conducted a simulation study to compare mean dose-response function estimate and corresponding 95% confidence band by our method with those from bootstrap method. We show that our method provides a better estimate of the meta-analysis summary dose-response relationship. We also analyze data from the Quantitative Analyses of Normal Tissue Effects in the Clinic (QUANTEC) initiative.

EMAIL: ireis@miami.edu

3j. Two-Dimensional Distributed Lag Model

Yin-Hsiu Chen*, University of Michigan
Bhramar Mukherjee, University of Michigan

We consider a problem to associate a time-series measured health outcome (e.g. mortality counts) with two time-series measured exposures (e.g. air particulate matter levels and ozone levels) with possible two-way interaction between exposures. Constrained distributed lag model (CDLM) is popular with bypassing the collinearity problem through characterizing the regression coefficient as a function of lag and the lag effect estimates can be more efficient. Tukey's one degree-of-freedom model for non-additivity is motivated from

a latent model framework and is powerful in capturing the two-way interaction due to its parsimony. The both models lead to biased estimators when they depart from the underlying truth. We propose a Bayesian constrained distributed lag model (BCDLM) to allow shrinkage between the saturated model and CDLM in order to reach an optimal bias-variance trade-off. We present the corresponding penalized likelihood approach and demonstrate that lower asymptotic mean squared errors (MSE) can be achieved in our framework. We use the data from National Morbidity, Mortality, and Air Pollution Study (NMMAPS) to illustrate our method.

EMAIL: yinhsiuc@umich.edu

3k. Interval Estimation of Ratio of Two Within-Subject Coefficients of Variation (WSCV) for Reproducibility Measure

Deukwoo Kwon*, University of Miami
Jeesun Jung, National Institute on Alcohol Abuse and Alcoholism, National Institutes of Health

In medical diagnosis and assay experiment, obtaining reliable and reproducible measurements is very important matter. The coefficient of variation (CV), intra-class correlation (ICC), and within-subject coefficient of variation (WSCV) can be used as an index of reliability/reproducibility of measurement. When the main focus is reproducibility, WSCV is a more appropriate measure since reproducibility is determined by how close repeated measurements within the same subject are. In this study, we develop methods for interval estimation of ratio of two WSCVs using the Wald-Type method, Fieller-Type method, log method, and method of variance estimates recovery (MOVER). In addition, we also propose a Bayesian method to obtain credible interval for ration of two WSCVs. Then we compare those frequentist methods with Bayesian method. We conduct a simulation study to illustrate empirical coverage rates and error rates of the proposed methods. For a real data example, we analyze the data from a study using the Computer-Aided Tomographic Scans (CAT-SCAN) in pediatric patients.

EMAIL: DKwon@med.miami.edu

3l. A Bayesian Joint Frailty-Copula Approach for Clustered Bivariate Time to Event Processes with Marked Data

Zheng Li*, Penn State College of Medicine
Ming Wang, Penn State College of Medicine
Vernon M. Chinchilli, Penn State College of Medicine

Copula is a popular approach to model bivariate time to event processes. However, when the corresponding bivariate outcomes are clustered, the traditional copula approach only accounts for the correlation between bivariate time to event processes but ignore the correlation within cluster. In order to account for both correlations between bivariate time to event processes and within cluster, we propose a joint frailty-copula approach under Bayesian framework to model clustered bivariate time to event processes. In addition, researchers may be interested in the effect of a variable ("marked data" measured only when events occur) on the time to event processes. In order to model such process, we incorporate the latent trait model into the joint frailty-copula model. The proposed method can be applied to analyze clustered bivariate time to event processes such as clustered competing risks, semi-competing risks, meta- analysis and etc. Extensive simulation studies are preformed to evaluate the performance of our method with respect to bias and mean squared error.

EMAIL: zhengli@hmc.psu.edu

3m. Bayesian Hierarchical Modeling of Substate Area Estimates from the Medicare CAHPS Survey

Tianyi Cai*, Harvard University

Each year, surveys are conducted to assess the quality of care for Medicare beneficiaries, using instruments from the Consumer Assessment of Healthcare Providers and Systems program. Depending on the heterogeneity of survey measures for Fee-for-Service beneficiaries in each state, the results are currently presented pooled at the state level or unpooled for substate areas. We fit spatial-temporal Bayesian random effects models using a generalized specification to estimate mean scores for each of the domains formed by 94 substate areas in 32 states over 5 years. A Bayesian hat matrix

provides a heuristic interpretation of the way the model combines information for estimates in these domains. The model can be used to choose between reporting of state or substate level direct estimates in each state, or as a source of alternative small area estimates superior to either direct estimate.

EMAIL: cai01@fas.harvard.edu

4. POSTERS: Imaging

4a. Double-Wavelet Transform for Multi-Subject Task-Induced Functional Magnetic Resonance Imaging Data

Minchun Zhou*, Vanderbilt University
Hakmook Kang, Vanderbilt University
David Badre, Brown University

The goal of this study is to model multi-subject task-induced fMRI response among predefined regions of interest (ROIs) of the human brain. Conventional approaches to fMRI analysis only take into account temporal correlations but do not rigorously model the underlying spatial correlation due to the complexity of estimating and inverting the spatio-temporal covariance matrix. Other spatio-temporal model approaches usually estimate the covariance matrix, which requires the signal to be stationary. To address these limitations, we propose a single level double-wavelet approach that transforms the model and data twice using different wavelet functions, where the wavelet coefficients are used to estimate model parameters and test hypotheses. Working with the wavelet coefficients simplifies temporal and spatial covariance structure because the wavelet coefficients are approximately uncorrelated. The wavelet transform is well-equipped to handle the non-stationary signals. Simulation studies showed that ignoring spatial correlation caused higher false positive and false negative error rates. We also applied our method to fMRI data to study activation in the prefontal cortex.

EMAIL: minchun.zhou@vanderbilt.edu

4b. Playing with Pie Chart in R: Filling the Slices with Black-and-White Patterns, Patterns with Colors, or Images

Chunqiao Luo*, University of Arkansas for Medical Sciences
Shasha Bai, University of Arkansas for Medical Sciences

Currently in R, pie charts generated using functions such as pie() in package graphics, pie3D() in package plotrix, or coord_polar() in package ggplot2 can only be color coded. There are no R functions that have the capacity of filling pie chart slices with patterns. The ppiechart package is a tool for creating aesthetically pleasing and informative pie charts in R. It can plot either black and white (BW) pie charts or pie charts in colors, with or without filled patterns. BW pie charts filled with patterns are useful for publications, especially when journals only accept BW figures. On the other hand, pie charts in colors with or without patterns are useful for online publishing, poster and PowerPoint presentations. People with color blindness can also benefit from the pattern feature of this package. The ppiechart allows the flexibility of a variety of combinations of patterns and colors to choose from. It also has the ability to fill the slices with any external images in png or jpeg formats. In summary, ppiechart allows the users to be as creative as they can while creating pie charts!

EMAIL: cluo@uams.edu

4c. TPRM: Tensor Partition Regression Models with Applications in Imaging Biomarker Detection

Michelle F. Miranda*, University of Texas MD Anderson Cancer Center
Hongtu Zhu, University of Texas MD Anderson Cancer Center
Joseph G. Ibrahim, University of North Carolina, Chapel Hill

Medical imaging studies have collected high dimensional imaging data in order to identify imaging biomarkers for diagnosis, screening, and prognosis, among many others. These imaging data are often represented in the form of a multi-dimensional array, called a tensor. The aim of this paper is to develop a tensor partition regression modeling (TPRM) framework to establish a relationship between low-dimensional clinical outcomes (e.g., diagnosis) and high dimensional tensor covariates. Our TPRM is a hierarchical model and efficiently integrates four components:(i) a parti-

tion model; (ii) a canonical polyadic decomposition model; (iii) a factor model; and (iv) a generalized linear model. This framework not only reduces ultra-high dimensionality to a manageable level, resulting in efficient estimation, but also optimizes prediction accuracy in the search for informative sub-tensors. Posterior computation proceeds via an efficient MCMC algorithm. We apply TPRM to predict Alzheimer's Disease from the structural magnetic resonance imaging data obtained from the ADNI study.

EMAIL: michellemirandaest@gmail.com

4d. Neuroconductor: A Framework for a Framework for Reproducible Neuroimaging Analysis in R

John Muschelli*, Johns Hopkins University
Jean-Philippe Fortin, University of Pennsylvania
Adrian Gherman, Johns Hopkins University
Brian Caffo, Johns Hopkins University
Ciprian M. Crainiceanu, Johns Hopkins University

Bioconductor is an extensive repository of R packages for bioinformatics, which has been used for over 10 years. It provides tutorials and courses to bioinformatics researchers to learn analysis. Moreover, it provides support for developers, data to test procedures, and a rigorous testing framework for packages. As statistics in neuroimaging analyses have grown in the past 5 years, we believe that we a need for a similar framework for neuroimaging. This framework will allow R users, particularly statisticians, to more easily learn how to perform neuroimaging analyses. Neuroconductor can leverage all the tools that R has to offer, such as state-of-the-art statistical tools, high-level package development, and reproducible tools like R Markdown and knitr. We have begun providing tutorials on how to perform basic imaging operations and analyses (http://johnmuschelli.com/neuroc/). The entire system is built on GitHub (https://github.com/neuroconductor-devel) where packages are uploaded and automatically checked using the Travis continuous integration system (http://travis-ci.org). We hope to enable more statisticians to break into neuroimaging with less effort.

EMAIL: jmuschel@jhsph.edu

4e. Bayesian Ensemble Models for Prostate Cancer Pattern Detection

Anjishnu Banerjee*, Medical College of Wisconsin
Tucker Keuter, Medical College of Wisconsin
Peter S. LaViolette, Children's Hospital of Wisconsin and Medical College of Wisconsin
Amy L. Kaczmarowski, Medical College of Wisconsin
Sarah L. Hurrell, Medical College of Wisconsin

One in six men will be diagnosed with prostate cancer, improving diagnostic accuracy is essential for preventing unnecessary procedures. Current imaging techniques lack the ability to accurately diagnose prostate cancer grade, especially in peripheral and high fluid density regions. Multiparametric MRI is gaining acceptance as the standard of care for prostate imaging. We propose a new algorithm to complement and improve diagnosis by predicting underlying histological information based on prior radiographic and pathologic knowledge. This will give radiologists the ability to accurately pinpoint the location and grade of prostate cancer nodules, thereby improving quality of life the patients. In this proposal, we propose an "ensemble" classifier, which combines a set of classifiers, gaining prediction accuracy from each one. We develop a Bayesian nonparametric weighting scheme, which besides borrowing strength across classifiers, also helps mitigate issues due to minor misalignment between MR images and the histology. This possible gain in prediction accuracy is illustrated through real and simulated data examples.

EMAIL: anjishnu@gmail.com

4f. Detection of Prostate Cancer with Multi-Parametric MR Imaging Models under a Bayesian Probability Framework

Jin Jin*, University of Minnesota
Joseph Koopmeiners, University of Minnesota
Lin Zhang, University of Minnesota
Greg Metzger, University of Minnesota
Ethan Leng, University of Minnesota

Multi-parametric magnetic resonance imaging has recently received substantial attention as a non-invasive method for the

detection of prostate cancer. However, its application in clinical practice has been limited because diagnostic accuracy is subject to substantial human error. Previously, our group developed a quantitative, user-independent, voxel-wise classifier that outperformed any single imaging modality for the detection of prostate cancer, which utilized a naive approach that ignored the unique structure of the prostate. In this talk, we propose a novel classifier for prostate cancer detection under a Bayesian probability framework, which accounts for the unique structure of the prostate. In addition, we combine our classifier with a spatial smoother to account for spatial correlation in the data. Our results show that the proposed classifier achieves significant improvement in prostate cancer detection compared to the naive model that does not account for the structure of the prostate. Its Bayesian structure also has the advantage of dealing with missing data, which is important for the practical application of our model.

EMAIL: jinxx493@umn.edu

4g. Bayesian Modeling of Medical Imaging in Tumor Delineation

Nitai D. Mukhopadhyay*, Virginia Commonwealth University
Kingston Kang, Virginia Commonwealth University

Medical images from dependent and independent sources are used in tumor delineation for cancer treatment. However, analysis of these images is rather difficult due to their massive dimensions. Therefore, performing any statistical analysis is possible only after reducing the image data to a small dimensional derivative of the images. Crainiceanu et al (2011) proposed population value decomposition (PVD), which can effectively reduce the size of a population of two-dimensional images. PVD reduces the image matrix to population specific matrix and much smaller dimensional subject specific matrix. Population-specific matrices do not change across different images, and subject-specific matrix is unique to each individual image. We use PVD to reduce the dimension of two-dimensional images and make a model for the low-dimensional subject-specific matrices. Bayesian modeling of the lower dimensional matrices provides a way to denoise the images and make inference based on the actual image data and provide model based predictions.

EMAIL: nitai.mukhopadhyay@vcuhealth.org

4h. Association of Structural Brain Imaging Measures with HIV Markers Incorporating Structural Connectivity Information: A Regularized Statistical Approach

Marta Karas*, Indiana University, Bloomington
Damian Brzyski, Indiana University, Bloomington
Beau Ances, Washington University in St. Louis
Timothy W. Randolph, Fred Hutchinson Cancer Research Center
Jaroslaw Harezlak, Indiana University, Bloomington

Brain imaging studies collect multiple imaging data types, but most analyses are done for each modality separately. Statistical methods that simultaneously utilize and combine multiple data types can instead provide a more holistic view of brain function. We develop a regularized statistical approach utilizing both structural information and connectivity. We estimate the model parameters by a unified approach directly incorporating structural connectivity information into the estimation by exploiting the joint eigenproperties of the predictors and the penalty operator. We apply this method to model associations between HIV markers (CD4 count and HIV RNA level) and cortical thickness and integrated rectified mean curvature measures obtained by FreeSurfer software while incorporating prior information from the structural connectivity between cortical regions.

EMAIL: j.harezlak@gmail.com

4i. A Bayesian Double Fusion Model for Resting State Brain Connectivity Using Joint Functional and Structural Data

Hakmook Kang*, Vanderbilt University
Hernando Ombao, University of California, Irvine
Christopher Fonnesbeck, Vanderbilt University
Zhaohua Ding, Vanderbilt University
Victoria L. Morgan, Vanderbilt University

Current approaches separately analyze concurrently acquired diffusion tensor imaging (DTI) and functional magnetic resonance imaging (fMRI) data. The primary limitation of these approaches is that they do not take advantage of the information from DTI that could potentially enhance estimation of resting state functional connectivity (FC) between brain regions. To overcome this limitation, we develop a Bayesian hierarchical spatio-temporal model

that incorporates structural connectivity into estimating FC. In our proposed approach, structural connectivity (SC) based on DTI data is used to construct an informative prior for functional connectivity based on resting state fMRI data via the Cholesky decomposition. Simulation studies showed that incorporating the two-data produced significantly reduced mean squared errors compared to the standard approach of separately analyzing the two data from different modalities. We applied our model to analyze the resting state DTI and fMRI data collected to estimate FC between the brain regions that were hypothetically important in the origination and spread of temporal lobe epilepsy seizures.

EMAIL: hakmook.kang@vanderbilt.edu

4j. PREVAIL: Predicting Recovery through Estimation and Visualization of Active and Incident Lesions

Jordan D. Dworkin*, University of Pennsylvania
Elizabeth M. Sweeney, Rice University
Matthew K. Schindler, National Institute of Neurological Disease and Stroke, National Institutes of Health
Salim Chahin, University of Pennsylvania
Daniel S. Reich, National Institute of Neurological Disease and Stroke, National Institutes of Health
Russell T. Shinohara, University of Pennsylvania

We develop a model that integrates imaging and clinical information at lesion incidence for predicting the recovery of white matter lesions in multiple sclerosis (MS) patients. Demographic, clinical, and magnetic resonance imaging (MRI) data were obtained from 60 subjects with MS at the National Institutes of Health. Imaging features were extracted from T1-weighted (T1w), T2-weighted, and magnetization transfer ratio (MTR) sequences acquired at lesion incidence. T1w and MTR signatures were also extracted from images one-year post-incidence. Baseline imaging features, clinical data, and demographic information were used to create statistical prediction models for long-term lesion damage. The root-mean-square prediction error was 0.95 for T1w and 0.064 for MTR, compared to measurement errors of 0.48 and 0.078. Three board-certified MS clinicians rated 100 lesions and found that predictions closely resembled true one-year appearance. This study shows that by using information from one visit at incidence, we can predict how a new lesion will recover. The potential to visualize the likely course of recovery has implications for clinical decision-making and trial enrichment.

EMAIL: jdwor@mail.med.upenn.edu

5. POSTERS: Diagnostics and Risk Factor Identification

5a. Comparison of Two Correlated ROC Surfaces at a Given Pair of True Classification Rates

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

The receiver operating characteristics (ROC) curve is typically employed when one wants to evaluate the discriminatory capability of a continuous biomarker in the case where two groups are to be distinguished, commonly the healthy and the diseased. There are cases for which the disease status has three categories. The ROC surface is a natural generalization of the ROC curve for three classes. In this paper, we explore new methodologies for comparing two continuous biomarkers that refer to a trichotomous disease status, when both markers are applied to the same patients. Comparisons based on the volume under the surface have been proposed, but that measure is often not clinically relevant. Here, we focus on comparing two correlated ROC surfaces at given pairs of true classification rates, which are more relevant to patients and physicians. We propose delta-based parametric techniques, power transformations to normality, and bootstrap-based smooth nonparametric techniques to investigate the performance of an appropriate statistic. We evaluate our approaches through an extensive simulation study and apply them to a real data set from prostate cancer screening.

EMAIL: leobantis@gmail.com

5b. A Case Study in Evaluating Diagnostic Tests Without a Gold Standard: From Kappa to Bayesian Latent Class Modeling

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

Operating characteristics of diagnostic tests are calculated based on knowledge of the test result and the true underlying disease status. However, a true gold standard test for the disease under study often does not exist, which leaves uncertainty about the true underlying disease status of the patient. One common approach is to have an expert panel examine the records of each patient to determine their disease status, providing an imperfect gold standard. The test(s) of interest are then compared for agreement with the expert panel and characteristics are estimated using the decision of the panel as the truth. Due to potential misclassification, this approach can introduce bias in either direction into the estimates. An alternative is to use Bayesian latent class models to estimate the characteristics of the test(s) while acknowledging the uncertainty surrounding the true disease status. Through a case study, we examine estimates of the characteristics using the Bayesian latent class model and the expert panel approach. We also discuss how to communicate the use of latent class models to clinician collaborators and solicit input for informative prior distributions.

EMAIL: kline.273@osu.edu

5c. Beta Regression for Modeling the ROC as a Function of Continuous Covariates

Sarah E. Stanley*, Baylor University
Jack D. Tubbs, Baylor University

The receiver operating characteristic (ROC) curve is a well-accepted measure of accuracy for diagnostic tests. In many applications, test performance is affected by covariates which should be accounted for in analysis. Several regression methods have been developed to model the ROC as a function of covariates within a generalized linear model framework. Two such methods, a parametric and semiparametric approach, estimate the ROC using binary indicators based on placement values that quantify the probability that a test result from a non-diseased subject exceeds that of a diseased subject with the same covariate values. A consequence of using binary indicators in this way is added correlation. As an alternative, we propose a new method that models the distribution of the placement values directly through beta regression. Given that the placement values are independent probabilities, a direct beta model is easily implemented and avoids the added correlation in the previous methods. We compare our beta method with the pre-existing parametric and semiparametric approaches via simulation and show that the new method yields comparable ROC estimates without adding correlation.

EMAIL: sarah_stanley@baylor.edu

5d. Statistical Methodology Survey Procedure using Published Medical Research

Noriko Tanaka*, National Center for Global Health and Medicine
Junya Shitaoka, Waseda University
Hayato Yamana, Waseda University

It is known that there is variation regarding the common statistical methods and terms as used in each medical research field. Extracting statistical methods described in medical articles is useful for statisticians not only for understanding the characteristic and trend in a medical field to conduct meta-analyses and systematic reviews or to conduct statistical consulting, but also for getting a new idea to develop statistical methodologies. Especially it is hard to keep up with the ever-progressing developments of statistical methodologies in bioinformatics. We present the statistical methodology survey procedure utilizing the literatures from the PubMed Central Open-Access archive. Simple application of text mining approach is not accurate to describe characteristics especially of the observational studies with tissue samples because the manuscript writing is not strictly regulated like clinical trials. So, we first described the difficulties and problems in collecting the information and present our procedure to show the characteristics of methodologies used in human brain tissue research.

EMAIL: denchu69@gmail.com

5e. Optimal Pool Sizes for Group Testing

Christopher R. Bilder*, University of Nebraska, Lincoln
Joshua M. Tebbs, University of South Carolina
Christopher S. McMahan, Clemson University
Brianna D. Hitt, University of Nebraska, Lincoln

Group testing is an indispensable tool for laboratories when testing high volumes of clinical specimens for infectious diseases. An important decision that needs to be made prior to its implementation is the group (pool) sizes to use. In best practice, an objective function is chosen and then minimized to determine an optimal set of group sizes. There are a few options for these objective functions, and they differ based on how the expected number of tests, assay characteristics, and laboratory constraints are taken into account. The purpose of this presentation is to closely examine a few common objective functions. We show the group sizes and/or results from using these objective functions are largely the same for standard testing algorithms in a wide variety of situations.

EMAIL: chris@chrisbilder.com

5f. A Novel Bayes Approach for Predicting Breast Cancer Risk Given Family History, with Application to the NIEHS Sister Study

Yue Jiang*, University of North Carolina, Chapel Hill
Clarice R. Weinberg, National Institute of Environmental Health Sciences, National Institutes of Health
Dale P. Sandler, National Institute of Environmental Health Sciences, National Institutes of Health
Shanshan Zhao, National Institute of Environmental Health Sciences, National Institutes of Health

Background: Breast cancer is a leading cause of cancer morbidity and mortality among U.S. women, with family history as a major risk factor. In the widely-used Gail model, family history is trichotomized as a 0/1/1+ variable. However, this paradigm loses information regarding the impact of family history on risk. We develop a well-calibrated family history score that improves predictive power by utilizing more information. Methods: We derive a Bayesian estimator of family-specific lifetime breast cancer risk that incorporates prior data from SEER registries and is updated with observed family structure, breast cancer history of first-degree female relatives, and cumulative hazard experienced among such relatives. We further calibrate our score with known risk factors such as age at menarche, age at first live birth, menopause, and others, through the Cox proportional hazards model, and evaluate the performance through receiver operation characteristic (ROC) curve analysis. Results: In the Sister Study cohort, our Bayesian score shows good calibration and increases predictive power compared to the Gail model in terms of estimated 5-year risk.

EMAIL: yuejiang@live.unc.edu

5g. Socioeconomic and Health-Related Risk Factors and Their Relative Importance for Health-Related Quality of Life

Ashley P. Fowler*, North Carolina A&T State University
Seong-Tae Kim, North Carolina A&T State University

Personal perception of Health-Related Quality of Life (HRQOL) is important because it is a predictor of morbidity, mortality, and needs for health care services. This study aimed to identify socioeconomic and health-related factors associated with HRQOL and their relative importance. We considered four self-reported HRQOL variables - overall health status, physical health status, mental health status, and activity limitations - with socioeconomic and health-related factors from the 2014 BRFSS data. We analyzed logistic regression along with subset and relative weight analysis. Overall and physical health status and activity limitations were most significantly affected by employment status, income, internet accessibility, arthritis, depression, and exercise. Mental health status was highly associated with smoking, alcohol, and marital status in addition to depressive disorder. This study identified statistically significant socioeconomic and health-related factors associated with HRQOL variables, which would contribute to potential guidelines for health promotion activities.

EMAIL: skim@ncat.edu

5h. Causal Modeling of Signaling Drivers in Head and Neck Cancer

Elizabeth Eisenhauer*, The College of New Jersey
Paloma Hauser, University of North Carolina, Chapel Hill
Michael F. Ochs, The College of New Jersey

Significant effort has been applied on gathering and analyzing data for different cancers to identify drivers of tumorigenesis. Most studies remain correlative, often due to a focus on identifying new pathways or molecules. We focused instead on identifying causal relationships from molecular data in head and neck cancer (HNC) assuming that receptors, pathways, and downstream regulators were known. Using two published data sets, we created a structural equation model (SEM) linking five receptors associated with HNC, four key signaling pathways (RAS-RAF, PI3K, JAK-STAT, and Notch), and six downstream transcriptional regulators. Using outlier analysis of copy number and methylation data on receptors, we estimated a probability of receptor activity in each sample. We used patient-specific gene sets of transcriptional targets to estimate a probability of transcriptional regulator activity for each sample as well. Both of these methods have been published previously by our group. We then fit the SEM model using these probability estimates. Our results showed causal relationships between EGFR and the JAK-STAT pathway, and the cMET and RAS-RAF pathway.

EMAIL: ochsm@tcnj.edu

5i. Size Investing Strategy on Multiple Confidence Intervals under FWER

Taeho Kim*, University of South Carolina
Edsel Pena, University of South Carolina

For confidence interval (CI) problems, researchers generally attempt to minimize the length of the interval, maintaining the coverage probability. By extending this approach to multiple CIs, the optimal Size Investing Strategy is investigated given the global coverage probability 1-q (equivalently global size q) under FWER in CI sense. To do this, a loss function approach for CI is adopted. The optimal size investing strategy for multiple CIs suggests to match the smaller confidence coefficients for larger standard errors to compensate the effect of the standard errors on the total length. This is different from the optimal size investing strategy for multiple testings (Pena et al., 2011&2015) because the trade-off relation between coverage probability and interval length in multiple CIs is different from the relation between size and power in multiple testings. Result shows about 5% of the total length reduction compared to the total length by Sidak procedure on the 1,000 location parameters of normal random variables.

EMAIL: taeho@email.sc.edu

5j. A Risk Stratification Approach for Improved Interpretation of Diagnostic Statistics

WITHDRAWN

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

We interpret standard diagnostic accuracy statistics, such as Youden index and AUC, in light of risk-stratification (how well a biomarker separates those at higher risk from those at lower risk) to better understand implications for public-health programs. We introduce an intuitive statistic, Mean Risk Stratification (MRS): the average change in risk (pre-test vs. post-test) revealed for tested individuals. MRS is a function of both AUC and disease prevalence, and thus little risk-stratification is possible for rare diseases (regardless of AUC), demonstrating a high-bar to justify population-based screening. AUC measures multiplicative relative gains in risk-stratification: AUC=0.6 achieves only 20% of maximum risk-stratification (AUC=0.9 achieves 80%). However, large relative gains in risk-stratification might not imply large absolute gains if disease is rare. We use MRS to compare cervical cancer screening tests in China vs. USA. The test with the worst AUC=0.72 in China (visual inspection with ascetic acid) provides twice the MRS of the test with best AUC=0.83 in the USA (HPV and Pap cotesting) because China has three times more cervical precancer/cancer.

EMAIL: katkih@mail.nih.gov

6. POSTERS: Clinical Trials and Biopharmaceutical Research Methods

6a. Design Considerations in Intervention Studies with Semicontinous Outcomes

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

The primary endpoints considered in many intervention studies are semicontinuous characterized by a mass at zero and continuous measurements (e.g., intervention to promote consumption of vegetable and fruit). However, convention design methods that ignore the semicontinuity of the outcomes are often used that can be substantially inefficient. In the paper, we discuss various issues in comparison of semicontinuous outcomes and develop efficient methods for power and sample size calculations.

EMAIL: liua@mail.nih.gov

6b. Statistical Considerations for Studies Evaluating the Efficacy of a Single Dose of the Human Papillomavirus (HPV) Vaccine

Joshua Sampson*, National Cancer Institute, National Institutes of Health
Aimee Kreimer, National Cancer Institute, National Institutes of Health
Allan Hildesheim, National Cancer Institute, National Institutes of Health
Mitchell Gail, National Cancer Institute, National Institutes of Health

The National Cancer Institute (NCI) and Gates Foundation are co-sponsoring a large 24,000 girl study to evaluate the efficacy of a one-dose regimen of the prophylactic HPV vaccine. The first component of the study is a four-year non-inferiority trial comparing 1- to 2-dose regimens of the two licensed vaccines. The second component is an observational study that estimates the vaccine efficacy (VE) of each regimen by comparing the HPV infection rates in the trial arms to those in a contemporaneous group of unvaccinated girls referred to as the "survey cohort." We will present an overview of the Statistical Analysis Plan for this study, highlighting the benefit of defining non-inferiority on the absolute risk scale when the number of expected events is near 0 and describing our method for estimating VE in the absence of a randomized control arm.

EMAIL: joshua.sampson@nih.gov

6c. Percentile Estimation Methods and Applications

Qi Xia*, Temple University
Yi Tsong, U.S. Food and Drug Administration
Yu-Ting Weng, U.S. Food and Drug Administration

Percentile is ubiquitous in statistics and plays a significant role in the day-to-day statistical application. Not only it can be applied to screening and confirmatory cut-point determination in immunogenicity assays but also the general percentile formulation enriches the statistical literature for mean comparison between the reference group and test group in bioequivalence or biosimilarity studies, with the analytical biosimilarity evaluation and scaled average bioequivalence as special cases. Shen et al. (2015) proposed and compared the exact based approach with some approximated approaches in one sample scenario for cut-point determination. However, the exact based approach has the issue of computational time complexity. In this poster, we explored more approximated approaches for percentile estimation such as Method of Variance Estimates Recovery (MOVER) based approaches and Modified Large Sample (MLS) approaches. All these approximated approaches are compared with the exact based approach in one or two sample scenarios. The applications and performance comparison for each approach are displayed with numerical results.

EMAIL: qixia@temple.edu

6d. Estimation of Dosage Frequency of Pre-Exposure Prophylaxis Needed to Protect Against HIV Infection

Claire F. Ruberman*, Johns Hopkins Bloomberg School of Public Health
Michael A. Rosenblum, Johns Hopkins Bloomberg School of Public Health
Jon A. Steingrimsson, Johns Hopkins Bloomberg School of Public Health
Craig W. Hendrix, Johns Hopkins University School of Medicine

Randomized controlled trials estimating the prevention effect of pre-exposure prophylaxis (PrEP) have shown varying effect sizes; this difference has been partially attributed to differences in medication adherence. Using targeted maximum likelihood estimation (TMLE), we estimate the effect of longitudinal drug concentration on protection against HIV infection based on three randomized PrEP trials: VOICE, Partners PrEP, and the Pre-exposure Prophylaxis Initiative (iPrEx). We focus on the effect on risk of HIV infection of setting drug concentration to be sustained above a certain threshold in preventing HIV infection. In contrast to previous analyses from the VOICE trials, our estimates from the VOICE study show a significant effect of treatment, which may be due to addressing sensitivity to time-varying confounders and the time dependent nature of the adherence process. We employ marginal structural models to further analyze the effect of different concentrations of tenofovir over time on risk of infection.

EMAIL: claireruberman@gmail.com

6e. A Bayesian Dose-Finding Design for Phase I/II Clinical Trials

Suyu Liu*, University of Texas MD Anderson Cancer Center

We propose a Bayesian phase I/II dose-finding trial design that simultaneously accounts for toxicity and efficacy in heterogeneous patient population. In oncology clinical studies, phase I part focuses on a general patient population which may cover several different tumor types, while phase II part often focuses on a specific tumor type. We apply the hierarchical prior model to effectively use the toxicity and efficacy information from different patient populations, to identify the recommended dose for future studies. An intuitive utility function that reflects the desirability trade-offs between efficacy and toxicity is used to guide the dose assignment and selection. We conduct extensive simulation studies to examine the operating characteristics of the proposed method under various practical scenarios. The results show that the proposed design possesses good operating characteristics and is robust to the shape of the dose-toxicity and -efficacy curves.

EMAIL: syliu@mdanderson.org

6f. Development of Methods for Shelf-Life Determination with Large Number of Batches

Sungwoo Choi, U.S. Food and Drug Administration
Yu-Ting Weng, U.S. Food and Drug Administration
Crystal Yu Chen*, University of Wisconsin, Madison
Meiyu Shen, U.S. Food and Drug Administration
Yi Tsong, U.S. Food and Drug Administration

To determine the labeled shelf life, the ICH Stability guidelines require at least three batches to be tested to allow for a reliable batch-to-batch variability. If the stability data passes the preliminary testing of batch similarity at the recommended significance level of 0.25, then a single expiration dating period can be estimated by pooling the stability data. However, if the stability data failed to pass the preliminary pooling test, an expiration dating period should be determined based on the minimum of the individual shelf lives. As of now, the preliminary pooling test and the minimal approach criteria for shelf life determination of more than three batches have not been addressed by the current ICH Stability guidance. We wish to discuss potential approaches for the scenario as mentioned. For the preliminary pooling test, we implemented the equivalence based testing method and determined the equivalence margin to be used in order to maintain a certain rejection power. As for the minimum approach criteria, we used a binomial proportion confidence interval approach to determine which "less favorable" batch could be used for large number of batches.

EMAIL: cchen397@wisc.edu

6g. A Web Application for Optimal Selection of Adaptive Designs in Phase I Oncology Clinical Trials

Sheau-Chiann Chen*, Vanderbilt University Medical Center
Yu Shyr, Vanderbilt University Medical Center

In phase I oncology clinical trials, research has shown that in some instances, the traditional 3+3 design performs worse than the continual reassessment method (CRM) design or the modified toxicity probability interval (mTPI) design and so on. This points to a lack of comprehensive evaluation for adaptive designs with flexible options, such as desired parameters setting. Thus, an interactive web application has been developed to find an appropriate adaptive design for conducting a real trial. The web application evaluates 11 different designs: two 3+3 designs, accelerated titration design (ATD), biased coin design (BCD), k-in-a-row (KIR) design, two CRM designs, escalation with overdose control (EWOC) design, escalation based on toxicity interval design, mTPI design and Bayesian optimal interval design (BOIN). Through simulation studies with a matched sample size, a comprehensive score is used to evaluate the performance of selected adaptive designs with desired parameters. The web tool provides interactive graphical user interface that allows users to easily conduct simulation and to assess the best design for meeting the primary objective of the proposed trials.

EMAIL: sheau-chiann.chen@vanderbilt.edu

6h. Comparison of Several Approaches to Adjust Overall Survival (OS) for Extensive Crossovers in Placebo-Controlled Randomized Phase 3 Trials

Shogo Nomura*, National Cancer Center, Chiba, Japan
Tomohiro Shinozaki, University of Tokyo
Chikuma Hamada, Tokyo University of Science

In recent years, innovative oncology agents have been actively developed. In placebo-controlled randomized phase 3 trials, setting the OS as primary endpoint, extensive crossovers in placebo arm are often unavoidable because such agents had shown embarrassingly favorable tumor responses. The problem is that the power of intention-to-treat (ITT) analysis is seriously lost. To regain the lost power, we propose a Bayesian framework with power prior which integrates ITT-based results for real-world data into that for re-constructed phase 3 data using a causal parameter estimated from a rank preserving structural failure time model. Numerical studies, motivated by a real phase 3 trial, were performed. The following analyses were set as benchmark: ITT; per-protocol which removed switchers; on-treatment which censored switchers when switching; time-varying Cox regression analysis. The performance was compared using unweighted log-rank test (LRT), Harrington-Fleming test, and optimally-weighted LRT (Bowden, et al, Statistics in Medicine, 2015). We will show results of the numerical study and discuss the applicability of our proposed method.

EMAIL: shnomura@east.ncc.go.jp

6i. Addressing Treatment Contamination in Trials of Complex Interventions: Measuring and Accounting for it in the Analysis

Nicholas P. Magill*, King's College London
Khalida Ismail, King's College London
Paul McCrone, King's College London
Sabine Landau, King's College London

In mental health trials there is concern that the effect of treatment offer may be biased by the impact of contamination (spillover), where patients in the control arm receive the treatment. An earlier systematic review found that this commonly happens due to crossover of clinicians or communication between units in different trial arms. The aim of this research was to develop estimators for evaluating the effect of treatment amongst those who would comply with treatment offer. Such local average treatment effects are known as complier average causal effects in the context of treatment compliance. This work investigates how these methods can be applied to the problem of contamination and extends their use. The estimators were applied to the secondary analysis of a trial of nurse-delivered psychotherapy for people with poorly controlled diabetes. The ITT effect estimate showed little evidence of effectiveness. Treatment receipt amongst control patients was measured using two treatment fidelity scales. The efficacy estimate, which accounted for contamination by using these measures, was a little larger but not statistically significant.

EMAIL: nicholas.magill@kcl.ac.uk

6j. Adjustment for Categorization in Predictor Variables

Saptarshi Chatterjee*, Northern Illinois University
Shrabanti Chowdhury, Wayne State University
Sanjib Basu, University of Illinois, Chicago

Cutoff detection in prognostic variables is an important area of research in clinical data analysis. Medical practitioners often need to categorize predictor variables in order to interpret the association with the outcome in a more meaningful way. Maximally selected chi-square statistic is often used for categorizing predictor variables. However, it has the disadvantage of possibly inflating the type-1 error rate by a significant margin. Several adjustments to correct the p-value of this statistic have been proposed in literature, but remain less used in practice. In this thesis, we propose a permutation based cutpoint selection method to overcome the issue of multiple testing. We validate our findings through extensive simulation studies and illustrate that this method maintains appropriate type-1 error while providing good power.

EMAIL: schatterjee@niu.edu

6k. Noninferiority Studies with Multiple Reference Treatments

Li-Ching Huang*, Vanderbilt University Medical Center
Yu Shyr, Vanderbilt University Medical Center

Non-inferiority (NI) studies are gaining popularity in clinical trials. This is largely due to the fact that NI trials enable investigators to compare a new treatment to a reference treatment by verifying the former is not worse than the latter by more than a pre-specified, small amount (NI margin). The loss of efficacy in a new treatment can be compensated by other benefits, such as fewer side effects, lower costs, and/or simpler treatment regimens. In this study, we discuss a more complex structure in a NI trial, where multiple new treatments are simultaneously compared to multiple reference treatments. The methodology will be layout to handle the analysis with either homogeneous or heterogeneous variances among all treatment groups. Clinical examples provided for illustrative purposes.

EMAIL: li-ching.huang@vanderbilt.edu

6l. Safety Monitoring in a Pediatric Patient Choice Trial

Erinn M. Hade*, The Ohio State University
Peter Minneci, The Ohio State University
Soledad Fernandez, The Ohio State University

While safety monitoring procedures in randomized clinical trials are well established, similar methods have not been thoroughly investigated for patient choice trials (PCT). PCTs allow consented and eligible participants to choose treatment allocation.PCTs follow similar procedures as would a randomized clinical trial, however due to patient choice, analyses for monitoring and for final inference, need to account for this selection mechanism. Recent work has considered safety monitoring in the observational setting for vaccine trials and surveillance studies and has extended randomized trial group sequential methods to accommodate covariate adjustment. In a non-surveillance setting, we investigate the performance of the Lan-Demets error spending approach and the more recent, group sequential estimating equation approach, when adjustment for selection is made through inverse probability weighting.We describe these methods for trials of more modest size, with fewer interim looks, and for varying methods of estimation for the probability of treatment selection. These methods are applied to a surgical versus pharmacologic treatment trial for children.

EMAIL: hade.2@osu.edu

6m. Adaptive Prediction of Event Times in Clinical Trials

Yu Lan*, Southern Methodist University
Daniel F. Heitjan, Southern Methodist University and University of Texas Southwestern Medical Center

Interim analyses are often planned at landmark event counts in event-based clinical trials. Accurate prediction of these landmark dates is beneficial.Available methods to create such predictions include parametric cure and non-cure models and nonparametric approach. The parametric methods work well when their underlying assumptions are met, and the nonparametric method gives calibrated but inefficient predictions across a wide range of models. In the early stages of a trial, when predictions have the highest marginal value, it is difficult to infer the form of the underlying model, including

whether a cure fraction exists. We propose here an adaptive method that entertains predictions from a set of possible models, drawing predictions from the candidate model with the highest posterior probability. To capture the uncertainty in model selection, we apply a simulation strategy using Bayesian bootstrap. A Monte Carlo study demonstrates that the adaptive method produces prediction intervals that have good coverage and are slightly wider than non-adaptive intervals but narrower than nonparametric intervals. It leads to improved predictions with data from the CGD Study.

EMAIL: ylan@smu.edu

6n. Robust Methods for Improving Power in Group Sequential Randomized Trial Designs, by Leveraging Prognostic Baseline Variables and Short-Term Outcomes

Tianchen Qian*, Johns Hopkins University
Michael Rosenblum, Johns Hopkins University
Huitong Qiu, Johns Hopkins University

In group sequential designs, adjusting for baseline variables and short-term outcomes can lead to increased power and reduced sample size. We derive simple formulas for the efficiency gain from such variable adjustment using semiparametric estimators. The formulas reveal how the impact of prognostic variables is modified by the proportion number of pipeline participants, analysis timing, and enrollment rate. While strongly prognostic baseline variables are always valuable to adjust for, the added value from prognostic short-term outcomes can be quite limited. For example, the asymptotic equivalent sample size reduction from prognostic short-term outcomes is at most half of the reduction from an equally prognostic baseline variable if pipeline participants are at most 1/3 of total enrolled. The added value from prognostic short-term outcomes is generally smallest at later interim analyses which are the ones that tend to impact power the most. A practical implication is that in trial planning one should put priority on identifying prognostic baseline variables. Our results are corroborated by simulation studies based on data from a real trial.

EMAIL: qiantianchen.thu@gmail.com

7. POSTERS: Latent Variable, Clustering and Missing Data Methods

7a. INVITED POSTER: Propensity Score Weighting for Causal Inference with Multi-Stage Clustered Data

Shu Yang*, North Carolina State University

Propensity score weighting is a tool for causal inference to adjust for measured confounders. Survey data are often collected under complex sampling designs such as multistage cluster sampling, which presents challenges for propensity score modeling and estimation. In addition, for clustered data, there may also be unobserved cluster effects related to both the treatment and the outcome. When such unmeasured confounders exist, and are omitted in the propensity score model, the subsequent propensity score adjustment will be biased. We propose a calibrated propensity score weighting adjustment for multi-stage clustered data in the presence of unmeasured cluster-level confounders. The propensity score is calibrated to balance design-weighted covariate distributions and cluster effects between treatment groups. In particular, we consider a growing number of calibration constraints increasing with the number of clusters, which is necessary for removing asymptotic bias that is associated with the unobserved cluster-level confounders. We show that our estimator is robust in the sense that the estimator is consistent without correct specification of the propensity score model. We extend the results to the multiple treatments case. In simulation studies, we show that the proposed estimator is superior to other competitors. We estimate the effect of School Body Mass Index Screening on prevalence of overweight and obesity for elementary schools in Pennsylvania.

EMAIL: syang24@ncsu.edu

7b. Constrained Estimation of Source-Specific Air Pollution during On-Road Commutes

Jenna R. Krall*, George Mason University
Jeremy Sarnat, Emory University

Determining the amount of air pollution generated by a particular source, such as vehicle exhaust, is challenging

because observed pollution is frequently a mixture of pollutants generated by multiple sources. Source apportionment models are latent variable models that estimate source-specific pollution using observed concentrations of pollutants. When available, prior information can help to estimate source-specific pollution, however it is not always clear how to incorporate prior information in source apportionment. We measured in-vehicle pollution and lung function for scripted commutes in Atlanta, GA. We applied the Multilinear Engine (ME-2) to estimate source-specific pollution and constrained the source apportionment solution to incorporate prior information about sources of traffic pollution. The ME-2 approach estimated pollution from vehicle exhaust, brake-related, road dust, and secondary sources. Without additional constraints, it is often not possible to differentiate temporally correlated sources such as road dust and brake wear. Using estimated source-specific pollution, we found exposure to brake-related pollution was associated with decreased lung function.

EMAIL: jenna.krall@gmail.com

7c. Investigating Multiple Imputation in Cluster Randomized Trials

Brittney E. Bailey*, The Ohio State University
Rebecca R. Andridge, The Ohio State University
Abigail B. Shoben, The Ohio State University

Missing data in cluster randomized trials are often handled with parametric multiple imputation (MI), assuming multivariate normality and using random effects to incorporate clustering. Since data do not always satisfy this assumption, a nonparametric approach to MI is desirable. Predictive mean matching (PMM) is a nonparametric approach where missing outcomes are imputed with observed outcomes in the data from donors that are similar to the missing cases. It is not clear how best to extend PMM to multilevel data. Two possibilities are to ignore clustering in the imputation model or to include fixed effects for clusters. In parametric MI, ignoring clustering in the imputation model leads to underestimation of the MI variance, while including fixed effects for clusters tends to overestimate the variance. A mixed effects imputation model can be used as the basis for matching, but this is computationally intensive and increases reliance on distributional assumptions. To simplify computation and reduce bias in the estimated variance, we investigate a weighted PMM approach that incorporates both the fixed effects imputation model and the imputation model that ignores clustering.

EMAIL: bailey.857@osu.edu

7d. Missing Data in Canonical Correlation Analysis

Emily Slade*, Harvard University
Peter Kraft, Harvard University

Canonical correlation analysis (CCA) provides a global test and measure of association between two multivariate sets of variables measured on the same individuals, such as air pollutants and disease biomarkers. In these large multivariate settings, the proportion of subjects missing data on at least one variable can be high. Before performing CCA in practice, missing data has typically been handled by complete case analysis, unconditional mean imputation, or k-nearest neighbor's approaches. For each of these methods as well as more sophisticated imputation methods, we examine bias of the first canonical correlation and power of a test of association between the two sets of variables. Bias of the first canonical correlation is strongly linked to sample size. Thus, limiting the sample size in a complete case analysis can lead to strongly biased results even when the data are MCAR, a result different from standard regression analyses. All imputation methods performed similarly; thus, we recommend that researchers should impute missing data in CCA, but there is not much to gain by using a sophisticated method such as tree-based imputation over a simple method such as mean imputation.

EMAIL: eslade@fas.harvard.edu

7e. Impact of Rotation Methods in Factor Analysis for Identification of Postoperative Cognitive Deficit

Mary Cooter*, Duke University Medical Center
Wenjing Qi, Duke University Medical Center
Jeffrey N. Browndyke, Duke University Medical Center
Mark F. Newman, Duke University Medical Center
Joseph P. Mathew, Duke University Medical Center
Yi-Ju Li, Duke University Medical Center

Assessment of global cognitive change is often conducted with a battery of tests repeated over time. Due to the number of tests

and the latent factors being assessed, dimensional reduction via rotated factor analysis is often employed. Commonly used orthogonal rotations result in uncorrelated factors. Given that oblique rotations can lead to correlated factors, we sought to evaluate implications of choosing one rotation method over another. Analyzing two datasets (cardiac surgery patients, N=1480), we explored differences in cognitive deficit (≥1SD drop in at least one factor) via Kappa statistics and McNemar's tests, and characterized patients with alternate classifications between the methods. Kappa between the two rotations was 0.70 and 0.77, and rate of cognitive deficit differed by 13.6% and 10.3% in each dataset. While there is fair agreement, the orthogonal rotation has heavier tailed distributions leading to higher rates of deficit. Given the correlation of the latent cognitive factors, and the sensitivity of orthogonal rotation to extreme test scores, we believe the oblique rotation is preferable.

EMAIL: mary.cooter@dm.duke.edu

7f. Hot Deck Imputation for Mixed Typed Datasets using Model Based Clustering

Sarbesh Raj Pandeya*, Georgia Southern University
Haresh Rochani, Georgia Southern University

Multiple imputation is a commonly used method when addressing the issue of missing values. Hot deck imputation is distinctively different than others to ensure closeness to true variance in estimating the regression coefficients as it involves the replacement of unobserved values by observed values in similar units or cells. These cells are determined in terms of the closeness of each observation using various distance measures. But most of the distance measures can only be applied to continuous variables. Thus, there is a distinct problem when there are categorical covariates in the dataset. We proposed for a model based clustering procedure that uses a parsimonious covariance structure of the latent variable, following a mixture of Gaussian distributions to generate the imputation cells of mixed type dataset (i.e. datasets with continuous and categorical variables). The results of the simulated data showed demonstrated lower variance compared to the complete cases in estimation of regression coefficients.

EMAIL: sp03459@georgiasouthern.edu

7g. Fast Computation of Smooth Nonparametric of Mixing Distributions via Kernel Mixtures

Nicholas Henderson*, Johns Hopkins University

We present a method for smooth estimation of an arbitrary mixing density which models the mixing density as a finite mixture of kernel functions with fixed points of support. By drawing on approaches used in mirror descent and accelerated gradient descent, we outline a fast, scalable algorithm for estimating the mixing proportions, and thus the corresponding mixing densities. We discuss several approaches for selecting the kernel bandwidths and examine their relative performance. Illustration of the method is developed through several simulation studies and an example involving RNA sequencing data.

EMAIL: nhender5@jhmi.edu

7h. Unifying and Generalizing Confounder Adjustment Methods that Use Negative Controls

David C. Gerard*, University of Chicago
Matthew Stephens, University of Chicago

Hidden confounding is a well-known problem in many fields, particularly large-scale gene expression studies. Recent proposals to use control genes - genes known to be unassociated with the covariates of interest have led to new methods to account for hidden confounding. Going by the moniker Removing Unwanted Variation (RUV), there are many versions: RUV1, RUV2, RUV4, RUVinv, RUVrinv, RUVfun. Recently, RUV4 was recast in a framework that unifies it to other confounder adjustment procedures. In this paper, we (1) recast RUV2 in this same framework, (2) prove conditions for which a procedure can be considered both RUV2 and RUV4, calling the resulting procedure RUV3, (3) introduce what we call RUV5, which can be considered a generalization of all versions of RUV, and (4) in the context of RUV5, present a principled and modular Bayesian approach to confounder adjustment that has superior performance and calibration to existing versions of RUV.

EMAIL: gerard.1787@gmail.com

7i. Application of Doubly-Robust Method for Missing Outcome Data to a Multilevel Longitudinal Survey

Nicole M. Butera*, University of North Carolina, Chapel Hill
Donglin Zeng, University of North Carolina, Chapel Hill
Annie Green Howard, University of North Carolina, Chapel Hill
Amy H. Herring, University of North Carolina, Chapel Hill
Penny Gordon-Larsen, University of North Carolina, Chapel Hill

The China Health and Nutrition Survey (CHNS) is an ongoing multilevel longitudinal cohort study, which like other longitudinal cohort studies, suffers from missing outcome data that can introduce bias in regression effect estimates. However, appropriately handling missing data in CHNS can be challenging due to the correlated multilevel data. To address this limitation, we developed a doubly-robust method to appropriately handle missing outcome data in an unbiased fashion under conditions of correct specification of: (1) the outcome model and (2) the probability of missingness model. This method used multilevel models for (1) the outcome model and (2) the probability of missingness model, adjusting for covariates at different levels (e.g., community, household, individual). We applied this method to CHNS data to estimate the effect of total physical activity (MET-hours) on body mass index (kg/m2). This method allowed the estimation of unbiased effect estimates in the presence of correlated missing outcome data using multilevel covariates.

EMAIL: butera@live.unc.edu

7j. Two-Phase Sample Selection to Improve Efficiency for Binary X and Y

Paul M. Imbriano*, University of Michigan
Trivellore E. Raghunathan, University of Michigan

A two-phase survey design is typically used to estimate the mean of a single outcome of interest,Y, which is expensive to obtain. If a surrogate variable X can be measured cheaper, then sampling can be done in two phases, the first phase takes a random sample from the population and measures X, and the second phase uses a subsample of the first to measure Y. When Y and X are binary, then stratified sampling is used to select the second phase sample. Ideally, we would choose a sample size for each stratum in order to minimize the variance of our estimate for the mean. Unfortunately, the variance depends on both P(Y = 1 | X = 0) and P(Y = 1 | X = 1), which may not be known beforehand. To overcome this limitation, we propose an adaptive sampling strategy, where the second phase sampling is done in batches. We also provide similar strategies for two-phase sampling when the risk difference and odds ratio are of primary interest. We tested the performance of our adaptive design against other methods for sample selection through simulations.

EMAIL: pimbri@umich.edu

7k. Leveraging Mixed and Incomplete Outcomes via a Generalized Reduced Rank Regression

Kun Chen, University of Connecticut
Chongliang Luo*, University of Connecticut
Jian Liang, Tsinghua University, China
Gen Li, Columbia University
Fei Wang, Cornell University
Changshui Zhang, Tsinghua University, China
Dipak Dey, University of Connecticut

Large-scale predictive modeling tasks routinely emerge in various fields. In many real-world problems, the collected outcomes are of mixed types, including continuous measurements, binary indicators and counts, and the data may also subject to substantial missing values. These mixed outcomes are often interrelated, representing diverse reflections or views of the same underlying data generation mechanism. We develop a mixed-outcome reduced-rank regression, which effectively and conveniently enables information sharing among all the prediction tasks. Our approach integrates mixed and partially observed outcomes belonging to the exponential dispersion family, by assuming that all the outcomes are associated through a shared low-dimensional subspace spanned by the high-dimensional features. A general regularized estimation criterion is proposed, and a unified algorithm with convergence guarantee is developed for optimization. Under a general sampling scheme of missing, we establish non-asymptotic performance bound for the proposed estimators. The effectiveness of our approach is demonstrated by simulation studies and an application of longitudinal studies of aging.

EMAIL: chongliang.luo@uconn.edu

8. POSTERS: Longitudinal and Survival

8a. Quantifying Practice Effects of Cognitive Assessment using the Penn "Computerized Neurocognitive Battery" (CNB) in a Neurodevelopmental Longitudinal Cohort

Angel Garcia de la Garza*, University of Pennsylvania
Kosha Ruparel, University of Pennsylvania
Kevin Seelaus, University of Pennsylvania
Allison Port, University of Pennsylvania
Chad Jackson, University of Pennsylvania
Tyler Moore, University of Pennsylvania
Raquel Gur, University of Pennsylvania
Warren Bilker, University of Pennsylvania
Ruben Gur, University of Pennsylvania

While cognitive performance is known to improve with age, its extent and rate vary especially during the early years. A longitudinal approach can help better understand development change in cognition. However, longitudinal cognitive testing suffers from practice effects that are known to distort longitudinal trends. We propose a method to quantify practice effects and separate them from age effects in a longitudinal context. The data is from a group of 512 subjects with longitudinal observations from the Philadelphia Neurodevelopmental Cohort. Each subject took the Penn Computerized Neurocognitive Battery (CNB). The proposed method uses a generalized additive regression model to approximate developmental non-linear trajectories of cognition using normative data. We quantify the expected practice effects using these developmental trajectories. Furthermore, we identify significant separable developmental and practice effects using a bootstrap resampling and determine covariates associated with these practice effects. We propose this method as a way to quantify practice effects in longitudinal cognitive testing in developmental samples.

EMAIL: angelgar@upenn.edu

8b. MCI Subtypes are Heterogeneous and Preceded by Different Risk Factors: A Study on Incident aMCI and Incident naMCI

Andrea R. Zammit*, Albert Einstein College of Medicine
Richard B. Lipton, Albert Einstein College of Medicine
Mindy J. Katz, Albert Einstein College of Medicine
Carol A. Derby, Albert Einstein College of Medicine
Charles B. Hall, Albert Einstein College of Medicine

Our objective was to identify risk factors predicting amnestic and non-amnestic mild cognitive impairment. Participants (n=1118) were from the Einstein Aging Study, a longitudinal study of cognitive aging and dementia. We conducted Cox regressions to identify MCI predictors from categories of cognition, physical function, affect, co-morbidites, and subjective complaints. Over a mean of 4 years of follow-up, 67 participants developed aMCI (mean age=78.9 years, female=65.7%) and 52 developed naMCI (mean age= 78.7 years, female = 73.1%). Risk factors of incident aMCI were history of stroke (HR = 2.20, 95%CI=1.02-4.77) and subjective memory complaints (HR = 2.96, 95%CI=1.42-6.17); delayed recall was protective (HR=0.85,95%CI=0.80-0.91). Risk factors for incident naMCI included poor performance on Trail Making Test B (HR=1.20, 95%CI=0.92-1.82), and higher scores on the Geriatric Depression Scale (HR=1.20, 95%CI = 0.70-0.95); protective factors included high Digit Span (HR=0.82, 95%CI=0.70-0.95) and better peak flow (HR= 0.76, 95%CI=0.61-0.94). Relevant to targeting prevention, our results indicate that aMCI and naMCI have distinct risk profiles.

EMAIL: andrea.zammit@einstein.yu.edu

8c. Statistical Monitoring of Clinical Trials with Semi-Competing Risks Outcomes

Toshimitsu Hamasaki*, National Cerebral and Cardiovascular Center
Scott R. Evans, Harvard School of Public Health
Tomoyuki Sugimoto, Kagoshima University
Koko Asakura, National Cerebral and Cardiovascular Center

Many clinical trials implement group-sequential designs. In some disease areas, e.g., oncology or cardiovascular disease,

these trials utilize event-time outcomes and are event-driven meaning that interim analyses are performed when a certain number of events have been observed. In some trials, it may be of interest to evaluate if a test intervention is superior to a control intervention on at least one of the event-time outcomes. In such trials, one challenge is how to monitor multiple event-time outcomes in a group-sequential setting as the information fraction for the outcomes may differ at any point in time. We discuss logrank test-based methods for monitoring two event-time outcomes in group-sequential trials that compare two interventions using two time-to-event outcomes. We evaluate two situations: (i) both events are non-composite but one event is fatal, and (ii) one event is composite but the other is fatal and non-composite. We consider several strategies for testing if a test intervention is superior to a control intervention on at least one of the event-time outcomes.

EMAIL: toshi.hamasaki@ncvc.go.jp

8d. Analysis of Longitudinal Competing Risk Data with Multiple Features

Tao Lu*, University of Nevada, Reno

Longitudinal competing risks data are often collected from clinical studies. Mixed-effects joint models are commonly used for the analysis of such data. Nevertheless, the following issues may arise in longitudinal survival data analysis: (i) most joint models assume a simple parametric mixed-effects model for longitudinal outcome, which may obscure the important relationship between response and covariates; (ii) clinical data often exhibits asymmetry so that symmetric assumption for model errors may lead to biased estimation of parameters; (iii) response may be missing and missingness may be informative. There is little work concerning all of these issues simultaneously. Motivated by an AIDS clinical data, we develop a Bayesian varying coefficient mixed-effects joint model with skewness and missingness to study the simultaneous influence of these features.

EMAIL: reno.lu@outlook.com

8e. An Approximate Joint Model for Multiple Paired Longitudinal Outcomes and Survival Data

Angelo F. Elmi*, The George Washington University
Katherine Grantz, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health
Paul S. Albert, National Cancer Institute, National Institutes of Health

Multivariate paired longitudinal data present interesting challenges to estimation in joint models due to the large dimensional random effects vector needed to capture correlation due to clustering with respect to pairs, subjects, and outcomes. Standard approaches based on the shared parameter model are too computationally intensive for most practical applications. We propose an alternative simpler approach where missing data is imputed based on the Posterior Predictive Distribution from a Conditional Linear Model (CLM) approximation. Existing methods for complete data are then implemented to obtain estimates of the survival parameters. Our method is applied to examine the effects of discordant growth in anthropometric measures of longitudinal fetal growth in twin fetuses and the timing of birth. Simulation results are presented to show that our method performs relatively well with small measurement errors under certain CLM approximations.

EMAIL: afelmi@gwu.edu

8f. Comparison of Survival Curves between Cox Proportional Hazards and Random Survival Forests in Survival Analysis

Brandon J. Weathers*, Utah State University
D. Richard Cutler, Utah State University

Survival analysis methods are a mainstay of the biomedical fields but are finding increasing use in other disciplines including finance and engineering. A widely-used tool in survival analysis is the Cox proportional hazards regression model. For this model, all the predicted survivor curves have the same basic shape, which may not be a good approximation to reality. In contrast the Random Survival Forests does not make the proportional hazards assumption and has the flexibility to model survivor curves that are of quite different shapes for different groups of subjects. We applied both techniques to a number of publicly available datasets

and compared the fit of the two techniques across the datasets using the concordance index and prediction error curves. In this process, we identified 'types of data' in which Random Survival Forests may be expected to outperform the Cox model.

EMAIL: brandon.weathers92@gmail.com

8g. Dynamic Prediction using Joint Models of Longitudinal and Recurrent Events Data: A Bayesian Perspective

Xuehan Ren*, University of Texas Health Science Center at Houston
Sheng Luo, University of Texas Health Science Center at Houston

In the clinical studies of cardiovascular diseases, multiple longitudinal variables (e.g., blood pressure, cholesterol) are measured and recurrent events (e.g., stoke and coronary heart disease) are recorded. Personalized prediction of the next occurrence of recurrent events is of great clinical interest because it enables physicians to make more informed decisions and recommendations for patients, leading to improved outcomes and increased benefit. We propose a joint model of longitudinal and recurrent event data. We develop a Bayesian approach for parameter estimation and a dynamic prediction framework for predicting target patients' future outcome trajectories and risk of next recurrent events, based on their data up to the prediction time point. Our method development is motivated by and applied to the Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT, n=42,418), the largest clinical study to compare the effectiveness of medications to treat hypertension.

EMAIL: xuehan.ren@uth.tmc.edu

8h. Modeling Progressive Disease Using Longitudinal Panel Data with Applications to Alzheimer's Disease Data

Jacquelyn E. Neal*, Vanderbilt University
Dandan Liu, Vanderbilt University

Alzheimer's disease (AD) is a slow, progressive disorder, with no fixed events that define its onset. Identifying older adults at highest risk of AD progression could benefit these patients through early interventions to prevent or delay the onset of AD. Panel data, or repeated measurements at pre-scheduled times, can be used to investigate transitions between disease stages. Logistic regressions on a cross-sectional subset of the data and transition models with a Markov assumption and baseline covariate values are the most common methods used when examining transitions between disease stages in AD. However, the assumptions of these models do not hold for the multistate nature of AD and its progressive nature. We have applied existing methodology from the literature to data from the National Alzheimer's Coordinating Center (NACC). We propose using partly conditional models to investigate transitions between disease stages, as these models are highly flexible, do not rely on the Markov assumption, and can incorporate time-dependent covariate information.

EMAIL: jacquelyn.e.neal@vanderbilt.edu

8i. Interval Censoring Ignorability in Cox Regression Models

Leah H. Suttner*, University of Pennsylvania
Sharon X. Xie, University of Pennsylvania

Despite the pervasiveness of interval censoring in clinical trials, there is little to no software available for handling this type of data. Instead, methods for right censoring, such as Cox regression models, are often used by imputing the event time as the right-endpoint of the censoring interval. To study the bias of this method, we use simulation studies considering a range of sample sizes, censoring interval lengths, and baseline hazards. We compare the bias, defined as the mean difference between the Cox model estimates and true parameters, when the hazard is modeled using true event times and right endpoints of the censoring interval. In small samples ($n<25$), the bias is large ($>10\%$) for all baseline hazards and interval lengths. In larger samples, the interval censoring bias increases slightly ($<10\%$) with interval length for constant and decreasing hazards, and increases greatly for increasing hazards. Additionally, we propose an index of interval censoring ignorability (ICI) to assess if it is sufficient to use right-endpoint imputation for a given study. We evaluate the performance of the ICI index based on its correlation with interval censoring bias.

EMAIL: lsutt@mail.med.upenn.edu

8j. Comparing the Tests for a Survival Curve at a Fixed Time Point in Single Arm Study

Isao Yokota*, Kyoto Prefectural University of Medicine

In single arm and relatively small size study like phase II rare cancer trials, clinicians are interested in x-year survival proportion. Cut-off probability as the null value is often determined and statistical testing are conducted. Such an analysis is called as milestone survival analysis (Chen TT, J Natl Cancer Inst. 2015), and will be one of options for endpoints. The difference between Kaplan-Meier estimates and cut-off value was tested using standard error from Greenwood's formula with some stabilizing transformations, such as log, logistic, complementary log-log and arcsine square root. While asymptotic characteristics may be same, these transformations affect type-I error rate and power in finite sample. Another approach to construct test statistics based on person-time method puts the null hypothesis that average rate up to a milestone time is equal to the value transformed from a cut-off probability to a rate scale. In this study, the testing performances due to choice of test statistics and stabilizing transformation are compared through Monte Carlo simulations. Also, the above methods are applied to real clinical trial example.

EMAIL: iyokota@koto.kpu-m.ac.jp

8k. Joint Model for Left-Censored Longitudinal Biomarker, Zero-Inflated Recurrent Events and Death Event in a Matched Study

Cong Xu*, Penn State College of Medicine
Ming Wang, Penn State College of Medicine
Vernon Chinchilli, Penn State College of Medicine

In many prospective cohort studies, researchers often invoke matching to control confounding effects and measure recurrent events during the follow-up. This event process can have a negative effect on health conditions, in which case the censoring time caused by death is likely to be informative. Longitudinal biomarkers are measured repeatedly over time, where there exists a left-censoring issue due to the inherent limit of detection. Thus, a joint model for a left-censored longitudinal biomarker, zero-inflated recurrent events and death is proposed under the Bayesian framework. We consider a matched logistic model for zero recurrent events. Incorporate one random effect to account for the correlation within biomarker measures as well as two frailties to measure the dependency between subjects within a matched pair and that among recurrent events within each individual. By sharing the random effects, death may be dependent on repeated biomarkers and recurrent event history. The posterior distribution is derived and computed for inference. Results are compared to maximum likelihood approaches via an MCEM algorithm and Gaussian quadrature. Extensive simulation studies are provided.

EMAIL: congxu@hmc.psu.edu

8l. Functional Joint Models for Longitudinal and Time-to-Event Data: An Application to Alzheimer's Disease

Kan Li*, University of Texas Health Science Center at Houston
Sheng Luo, University of Texas Health Science Center at Houston

Studies for Alzheimer's disease (AD) often collect repeated measurements of clinical variables, event history, and functional data, to better understand the disease. An accurate prediction of the time to dementia based on the information is particularly helpful for physicians to monitor patients' disease progression and make informative medical decisions. We propose a functional joint model (FJM) to account for functional predictors (high dimensional magnetic resonance imaging) in both longitudinal and survival submodels in the joint modeling framework. The FJM provides the accurate dynamic prediction of target patients' future outcome trajectories and risk of AD conversion, based on both scalar and functional measures. Our proposed model is evaluated by a simulation study and is applied to the Alzheimer's Disease Neuroimaging Initiative (ADNI) study, a motivating clinical study testing whether serial brain imaging, clinical and neuropsychological assessments can be combined to measure the progression of Alzheimer's disease.

EMAIL: kan.li@uth.tmc.edu

9. POSTERS: Time Series, Multivariate and Functional Data Analysis

9a. Using Optimal Test Assembly Methods to Shorten Patient-Reported Outcome Measures

Daphna Harel*, New York University
Alexander W. Levis, Lady Davis Institute for Medical Research and McGill University
Linda Kwakkenbos, Lady Davis Institute for Medical Research, McGill University and Radboud University
Brett D. Thombs, Lady Davis Institute for Medical Research and McGill University

Patient-reported outcomes (PROs) assess patient health, well-being, and treatment response based on patient perspectives. Inclusion of PRO measures has become central in many clinical trials and cohort-based observational studies. Thus, efficient measurement of PROs is essential to limit both cost of patient cohorts and burden to patients who may be asked to respond to many different scales. Shortening these measures while maintaining measurement equivalence could reduce burden without compromising data quality. Currently, there are no standard methods for creating shortened forms. Optimal test assembly (OTA), used widely in high-stakes educational test design, can automate the selection of a subset of items through mixed-integer programming by generating short forms that maintain maximum Fisher information across the continuum of the latent trait while satisfying user-given constraints, such as the shape of the test information function, or number of items per subscale. This study explores the applicability of OTA for shortening PRO measures through an example, the Cochin Hand Function Scale, commonly used in rheumatic diseases.

EMAIL: daphna.harel@nyu.edu

9b. Meta-Analyses with an Increasing Number of Parameters: Comparison of Multivariate and Univariate Approaches

Simina M. Boca*, Georgetown University Medical Center
Ruth M. Pfeiffer, National Cancer Institute, National Institutes of Health
Joshua N. Sampson, National Cancer Institute, National Institutes of Health

Meta-analysis averages estimates of multiple parameters across studies. Univariate meta-analysis (UVMA) considers each parameter individually, while multivariate meta-analysis (MVMA) considers the parameters jointly and accounts for the correlation between their estimates. We compare the performance of MVMA and UVMA as the number of parameters, p, increases. Specifically, we show that (i) for fixed-effect meta-analysis, the benefit from using MVMA can substantially increase as p increases; (ii) for random effects meta-analysis, the benefit from MVMA can increase as p increases, but the potential improvement is modest in the presence of high between-study variability and the actual improvement is further reduced by the need to estimate an increasingly large between study covariance matrix; and (iii) when there is little to no between study variability, the loss of efficiency due to choosing random effects MVMA over fixed-effect MVMA increases as p increases. We demonstrate these three features through theory, simulation, and a meta-analysis of risk factors for Non-Hodgkin Lymphoma.

EMAIL: smb310@georgetown.edu

9c. Bartlett Correction for Multivariate Random Effects Models in Network Meta-Analysis

Hisashi Noma*, The Institute of Statistical Mathematics

Network meta-analysis enables comprehensive synthesis of evidence concerning multiple treatments and their simultaneous comparisons based on both direct and indirect evidence. Although ordinary likelihood-based methods or Bayesian analyses with non-informative priors have been adopted for the inference in multivariate random effects models in network meta-analysis, validities of these methods are founded on large sample theory. As widely known in conventional pairwise meta-analyses, coverage probabilities of confidence intervals of these methods can be substantially below the target level (e.g., Brockwell and Gordon, Statist Med 2007, 26: 4531-43). One of effective approaches to resolve these inferences is adopting improved methods based on higher order asymptotic theory. In this study, I develop the Bartlett correction-based confidence interval for multivariate random effects models in network meta-analysis, and evaluate its practical

effectiveness via simulation studies. In addition, applications to a real data example is provided.

EMAIL: noma@ism.ac.jp

9d. Temporal Trends and Characteristics of Clinical Trials for which only One Racial or Ethnic Group is Eligible

Brian L. Egleston*, Fox Chase Cancer Center
Omar Pedraza, Ipsos Public Affairs
Yu-Ning Wong, Fox Chase Cancer Center
Candace L. Griffin, Johns Hopkins University
Eric A. Ross, Fox Chase Cancer Center
J. Robert Beck, Fox Chase Cancer Center

There has been increasing emphasis on ensuring diversity in clinical trials such that results are more generalizable to a broad population. However, diverse trials can increase heterogeneity in estimators, which reduces study power. This creates a classic bias-variance trade-off: Less diverse samples reduce variability at the expense of increasing bias with respect to the applicability of study findings to a wider group. We examined 19,199 trial eligibility requirements from ClinicalTrials.gov. We found that the odds of a study restricting eligibility to a single race or ethnicity has been increasing (Odds Ratio=1.04 per year, 95% confidence interval 1.01-1.08). After adjustment, industry sponsored trials, behavioral trials, and gender-specific trials were the most likely to require participants to be a single race or ethnicity. Such studies were more likely to open in ZIP codes with more residents of the same race or ethnicity. The pattern of eligibility restrictions might be due to efforts to reduce behavioral health disparities in minority communities, and to increase the development of racially targeted drugs.

EMAIL: brian.egleston@fccc.edu

9e. Canonical Correlation for Principal Components of Multivarite Time Series

S. Yaser Samadi*, Southern Illinois University
Lynne Billard, University of Georgia

With contemporary data collection capacity, data sets containing large numbers of different multivariate time series relating to a common entity (e.g., fMRI, financial stocks) are becoming more prevalent. One pervasive question is whether or not there are patterns or groups of series within the larger data set (e.g., disease patterns in brain scans, mining stocks may be internally similar but themselves may be distinct from banking stocks). We develop an exploratory data methodology which in addition to the time dependencies, utilizes the dependency information between S series themselves as well as the dependency information between p variables within the series simultaneously while still retaining the distinctiveness of the two types of variables. This is achieved by combining the principles of both canonical correlation analysis and principal component analysis for time series to obtain a new type of covariance/correlation matrix for a principal component analysis to produce a so-called "principal component time series". The results are illustrated on two data sets.

EMAIL: s.y.samadi@gmail.com

9f. P-splines with an L1 Penalty for Repeated Measures

Brian D. Segal*, University of Michigan

P-splines are a nonparametric regression method. In the typical formulation, P-splines use a B-spline basis, and enforce smoothness by applying a quadratic penalty to the discrete differences in B-spline coefficients. This is equivalent to adding an L2 penalty to the likelihood. Separate from P-splines, nonparametric regression methods that use an L1 penalty have gained interest, particularly l1 trend filtering, which is equivalent to locally adaptive regression splines in certain cases, and can be formulated as a generalized lasso. These L1 penalty methods have been shown to adapt to local differences in smoothness better than smoothing splines, which use a quadratic penalty. Motivated by these developments and the flexibility of P-splines, we propose L1 P-splines, in which the quadratic penalty is replaced with an L1 penalty. While L1 penalties have been used in conjunction with P-splines for quantile regression, to the best of our knowledge, L1 penalties have not been used with P-splines for mean regression. We investigate the advantages and disadvantages of L1 P-splines, focusing on applications to repeated measures data.

EMAIL: bdsegal@umich.edu

9g. Evaluating Particle Filtering Methods for Analyzing Time Series Data from Complex Multi-pathogen Disease Systems

Xi Meng*, University of Massachusetts, Amherst

Immunological interactions between dengue serotypes give rise to challenges in drawing inference for transmission parameters. A popular framework of likelihood-based inference, particle filtering, is among the few methods that could be applied in these situations. We applied this framework to an established SICR model for a 4-serotype dengue system to estimate interaction parameters. We used a tiered particle filtering approach across increasingly smaller two-dimensional grids of parameter space. This approach enabled us to characterize the likelihood surface of parameters of interest, and we tested it on a set of simulated data as well as real dengue case data. It took 2193 CPU hours to fit a single dataset containing 481 data points. In datasets simulated from realistic parameter values, we observed average bias of 16.5% and 54% in estimating antibody dependent enhancement and cross-protection duration, respectively. Additionally, the method showed varying degrees of sensitivity to misspecification in known parameters. Our analyses quantify the challenges of implementing a full-scale analysis using these methods for complex disease transmission models.

EMAIL: xmeng@schoolph.umass.edu

9h. Model Averaging for Probabilistic Time Series Forecasts

Evan L. Ray*, University of Massachusetts, Amherst
Nicholas G. Reich, University of Massachusetts, Amherst

Ensemble forecasts are widely considered to be state-of-the-art forecasting methods. In many real-time prediction settings, such as a public health agency monitoring infectious disease, analysts construct series of forecasts over time; for example, we might update our predictions each week as new data arrive. We develop an ensemble approach that combines probabilistic forecasts from multiple models through model averaging. The model weights are a function of observed inputs such as the time when the prediction is made and recent observations of the process being predicted. We estimate the model weights by optimizing the log-score of the combined predictive distribution subject to a penalty encouraging the weights to change slowly as a function of the observed covariates. We apply our method to influenza data using three component models: a seasonal autoregressive integrated moving average model, a semiparametric method combining kernel conditional density estimation and copulas, and a simple approach based on kernel density estimation. We demonstrate that averaging across all time points in a held-out test data set, the ensemble approach outperforms the individual component models.

EMAIL: evan.l.ray@gmail.com

9i. Using Regression Model to Analyze the Relationship Between Chinese Air Pollution and the Yield

Qinan Mao*, Wright State University

China is a country which is famous for its agriculture, especially for the planting. Since China started to implement reform and opening up policy, its economy has been growing up rapidly. However, this economic development is booming at the cost of sacrificing the environment, such as the air pollution of the particulate matter (PM2.5 and PM10). With the aid of SPSS, the statistical method and the relevant knowledge are used to test whether the relationship exists between the air pollution and the yield of planting. By using the data of recent years, a regression model will be constructed to examine the possible relationship.

EMAIL: mao.12@wright.edu

9j. Functional Data Analysis in Characterizing Bacterial Vaginosis Patterns and Identification of Pattern-Specific Risk Factors

Xiangrong Kong*, Johns Hopkins University
Rahul Biswas, University of Washington
Marie Thoma, University of Maryland

We introduce functional data analysis (FDA) for exploring frequently measured biomarker data for bacterial vaginosis (BV). BV is an abnormal vaginal condition and its natural history is not well characterized. Nugent scores for BV diagnosis was available

for 312 women from rural Uganda on a weekly basis for two years. We first applied functional principal components method for dimension reduction. Extending Ferraty et al's method, we then proposed an unsupervised classification method that uses multiple data features hierarchically for classification. Our analysis on the BV data identified four distinct patterns of BV scores that were not described in BV literature before: women with persistent BV over the two years; women with persistent low score; women with rapid fluctuation but more often with low score; and women with rapid fluctuation but more often with high BV score. To further our understanding of BV, we conducted further analysis using regression methods to identify pattern specific risk factors. This application demonstrates that FDA can be a useful method for characterizing infectious disease patterns.

EMAIL: xkong4@jhu.edu

9k. A Frequency Domain Approach to Stationary Subspace Analysis with Applications to Brain-Computer Interface

Raanju Sundararajan*, Texas A&M University
Mohsen Pourahmadi, Texas A&M University

Discovering stationary linear relationships among components of a multivariate time series provides useful insights in the analysis of various nonstationary systems in the real world. Stationary Subspace Analysis (SSA) is a technique that attempts to find those linear transformations of nonstationary processes that are stationary. This problem has been studied in the time domain where the nonstationarity is characterized in terms of means and lag-0 covariances of the observations. We propose a frequency domain method to find a stationary subspace of a multivariate second-order nonstationary time series. Using the asymptotic uncorrelatedness property of the discrete Fourier transform of a stationary time series, we construct a measure of departure from stationarity and optimize it to find the stationary subspace. The dimension of the subspace is found using a sequential testing procedure and the asymptotic consistency of this procedure is discussed. We illustrate the better performance of our frequency domain method in comparison to time domain based SSA methods through simulations and discuss an application in analyzing EEG data from Brain-Computer Interface experiments.

EMAIL: raanch316@tamu.edu

9l. Adaptive Bayesian Power Spectrum Analysis of Multivariate Nonstationary Time Series

Zeda Li*, Temple University
Robert Todd Krafty, University of Pittsburgh

This article introduces a nonparametric approach to multivariate time-varying power spectrum analysis. The procedure adaptively partitions a time series into an unknown number of approximately stationary segments, where some spectral components may remain unchanged across segments, allowing components to evolve differently over time. Local spectra within segments are fit through Whittle likelihood based penalized spline models of modified Cholesky components, which provide flexible nonparametric estimates that preserve positive definite structures of spectral matrices. The approach is formulated in a fully Bayesian framework, in which the number and location of partitions are random, and relies on reversible jump Markov chain Monte Carlo methods that can adapt to the unknown number of segments and parameters. By averaging over the distribution of partitions, the approach can capture both abrupt and slow-varying changes in spectral matrices. Empirical performance is evaluated in simulation studies and illustrated through an analysis of electroencephalography during sleep.

EMAIL: tuf27908@temple.edu

10. POSTERS: Longitudinal Data Methods

10a. Comparing Methods for Residual Time Modeling under an Illness-Death Model

Krithika Suresh*, University of Michigan
Jeremy M.G. Taylor, University of Michigan
Alexander Tsodikov, University of Michigan

Dynamic prediction incorporates time-dependent marker information accrued during follow-up to improve personalized survival prediction probabilities. At any follow-up (landmark) time, the residual time distribution for an individual, conditional on their updated marker values, can be used to produce a dynamic pre-

diction. To satisfy a consistency condition that links dynamic predictions at different time points (Jewell and Nielsen, 1993), the residual time distribution must follow from a prediction function that models the joint distribution of the marker process and time to failure, such a joint model. To circumvent modeling the marker process, the approximate method landmarking is used, which fits a Cox model at a sequence of landmark times. Considering an illness-death model, we derive the residual time distribution and demonstrate that the structure of the Cox model baseline hazard and covariate effects under the landmarking approach does not provide a good theoretical approximation. We compare the performance of landmark models with a multi-state model with Weibull transition intensities using simulation studies and cognitive aging data from the PAQUID study.

EMAIL: ksuresh@umich.edu

10b. Box-Cox Transformed Linear Mixed Model with AR(1) Correlation Structures in the Analysis of Neurosurgical Patient Blood Glucose Measurements

Zheyu Liu*, University of Texas Health Science Center at Houston
Dong H. Kim, University of Texas Health Science Center at Houston
Dejian Lai, University of Texas Health Science Center at Houston

In medical research, natural logarithmic transformation is a standard approach to achieve normality assumption of measurements, but an alternative transformation may yield a better strategy. The Box-Cox method helps to select an optimal transformation to ensure the validity of a Gaussian distribution in regression modeling. Instead of assuming that we know the transformation that makes the outcome normal and linearly related to the covariates, we estimate the transformation by using maximum likelihood approach within the Box-Cox family. We extended the Box-Cox transformation method to a linear mixed effects model with an autoregressive correlation structure. Analyses of longitudinal measurements of neurosurgical patient blood glucose with missing over time and a simulation study with different correlation structure will illustrate the benefits of the proposed methodology and the resulting proper transformation. Our study indicated that neurosurgical patient blood glucose measurements should be transformed to satisfy Gaussian distribution before making valid statistical inference.

EMAIL: zheyu.liu@uth.tmc.edu

10c. Marginalization of Regression Parameters from Mixed Models of Categorical Outcomes

Donald Hedeker*, University of Chicago
Stephen du Toit, Scientific Software International
Hakan Demirtas, University of Illinois, Chicago
Robert D. Gibbons, University of Chicago

This presentation focuses on marginalization of the regression parameters in mixed models for correlated categorical outcomes. The regression parameters in such models have the "subject-specific" (SS) interpretation, representing effects of the covariates conditional on the random effects. This is in contrast to the "population-averaged" (PA) or marginal estimates that represent the unconditional covariate effects, which can be obtained using generalized estimating equations (GEE) and marginalized multilevel models. For random-intercept models there is a simple relationship between the SS and PA estimates, however, for models with multiple random effects this is not the case. We describe an approach using numerical quadrature to obtain PA estimates from their SS counterparts in models with multiple random effects. Standard errors for the PA estimates are derived using the delta method. We illustrate our proposed method using data from a smoking cessation study in which a dichotomous outcome (smoking, Y/N) was measured longitudinally, and compare our estimates to those obtained using GEE and marginalized multilevel models.

EMAIL: hedeker@uchicago.edu

10d. Bayesian Nonparametric Multivariate Quantile Regression Models

Sungduk Kim*, National Cancer Institute, National Institutes of Health
Paul S. Albert, National Cancer Institute, National Institutes of Health

The appropriate interpretation of monitored fetal growth throughout pregnancy in individual fetus and population is

dependent on the availability of adequate standards. The focus of this paper is on developing Bayesian nonparametric multivariate quantile regression models to develop contemporary U.S. fetal growth standards for racial/ethnic groups of pregnant women. The proposed method relies on assuming the asymmetric Laplace distribution as auxiliary error distribution. We also consider the covariates-dependent random partition models that the probability of any particular partition is allowed to depend on covariates. This leads to random clustering models indexed by covariates, i.e., quantile regression models with the outcome being a partition of the experimental units. Markov chain Monte Carlo sampling is used to carry out Bayesian posterior computation. Several variations of the proposed model are considered and compared via the deviance information criterion. The proposed methodology is motivated by and applied to a longitudinal fetal growth study.

EMAIL: kims2@mail.nih.gov

10e. Testing Independence between Observations from a Single Network

Youjin Lee*, Johns Hopkins Bloomberg School of Public Health
Elizabeth Ogburn, Johns Hopkins Bloomberg School of Public Health

Subjects in a study are often recruited from a single or a small number of social networks, e.g. from a single community or hospital. When subjects share social network ties, observations sampled from them may not be statistically independent, possibly due to homophily or peer effects. Unfortunately, despite a surge of research on network effects, dependence among observations has often been obscured in the statistical models used to analyze network data, and independence has been taken for granted. Statistical inference based on the assumption of independence is likely to result in anti-conservative statistical inference. We define dependence that is informed by network topology as network dependence and propose a nonparametric test to assess whether network dependence is present. Our proposed test is a generalized version of spatial autocorrelation test, with flexible weighting for network-based distances. We show the validity of the test through simulated networks under plausible dependence schemes. The method is finally applied to network data from the Framingham Heart Study, which has been used for many celebrated network effect studies under independence assumption.

EMAIL: ylee160@jhu.edu

10f. Transition Models with Informative Cluster Size and Informative Gap Time: A Joint Modeling Approach

Joe Bible*, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health
Danping Liu, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health
Paul S. Albert, National Cancer Institute, National Institutes of Health

Transition models are useful in estimating the probability of recurrent events in longitudinal studies. However, direct application of a transition model may suffer from two complications, informative cluster size and informative gap time between observations. For example, Consecutive Pregnancy Study (CPS) is a retrospective cohort study aiming at understanding the recurrence patterns and predictors of adverse pregnancy outcomes, such as preterm birth. The number of pregnancies observed and the gap time may be both indicative of a women's underlying fertility, and hence correlated with the pregnancy outcomes. We propose a shared random effect structure for jointly modelling the transition model with the informative observation process. The gap time is modelled by a parametric distribution with right censoring; the cluster size is characterized by a continuation ratio model. We also investigated the estimation and interpretation of two transition probabilities: one adjusted for gap time and the other marginalized over gap time.

EMAIL: jbible831@gmail.com

10g. Predicting Gestational Age from Longitudinal Maternal Anthropometry when the Onset Time is Measured with Error

Ana Maria Ortega-Villa*, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health
Paul S. Albert, National Cancer Institute, National Institutes of Health

Determining the date of conception is important for estimating gestational age and monitoring whether the fetus and mother are on track in their development and pregnancy. Various methods based on ultrasound have been proposed for dating a pregnancy in high resource countries. However, such techniques may not be available in under resourced countries. We develop a shared random parameter model for estimating the date of conception using longitudinal assessment maternal anthropometry when the gold standard assessment of the onset time may be available in only a fraction of the subjects. The methodology is evaluated with a training-test set paradigm as well as with simulations to examine the robustness of the method to model misspecification. We illustrate this new methodology with data from the NICHD Fetal Growth Studies.

EMAIL: ana.ortega-villa@nih.gov

10h. Evaluation of Efficiency Gains and Bias in Joint Modeling of Outcomes of Different Types

Ralitza Gueorguieva*, Yale University
Eugenia Buta, Yale University
Denica Grigorova, Sofia University
Brian Pittman, Yale University
Stephanie O'Malley, Yale University

Joint modeling is often necessary to evaluate predictor effects on multiple outcomes, possibly repeatedly measured over time. In particular, correlated probit models are used for analysis of continuous and binary measures; mixture models are helpful for zero-inflated data; joint models of frequency and intensity measures evaluate different aspects of substance use behaviors. We consider novel models for joint analysis of categorical and continuous outcomes, maximum likelihood and Bayesian methods and software implementations. Simulations evaluate efficiency gains of simultaneous analysis compared to separate analysis. Potential bias under misspecification of the random effects structure will also be discussed. Data on frequency and intensity of drinking, and on the relationship between the use of traditional tobacco products and e-cigarettes illustrate the methods and discussion. Research is supported by Yale TCORS (P50DA036151) from NIDA, the FDA Center for Tobacco Products, and by CTNA (P50 AA-012870) grant from NIAAA. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or the FDA.

EMAIL: ralitza.gueorguieva@yale.edu

10i. Modeling Restricted Mean Survival Time Using Stratification

Xin Wang*, University of Michigan
Douglas Schaubel, University of Michigan

Restricted mean survival time (RMST) is often of great clinical interest in practice, and is gaining increased attention among biostatisticians. There are now several existing methods to model RMST, with the methods distinguished by their estimation approaches and assumptions on the censoring mechanism. We propose a stratified RMST model. The model assumes multiplicative covariate effects and applies to several practical settings. Examples include (i) clustered data and (ii) data with high-dimensional categorical covariate (e.g., center). The proposed methods are motivated by modeling of RMST among end-stage renal disease patients, in the presence of a high-dimensional covariate (dialysis facility). Estimation proceeds through a computationally efficient algorithm analogous to stratification. Large sample properties of the proposed estimators are derived and simulation studies are conducted to assess their finite sample performance. We apply the proposed methods to data obtained from the United States Renal Data System, which includes patients from over 6,000 dialysis facilities.

EMAIL: wangxinnju@gmail.com

10j. Regularity of a Renewal Process Estimated from Binary Data

John D. Rice*, University of Rochester
Robert L. Strawderman, University of Rochester
Brent A. Johnson, University of Rochester

Assessment of the regularity of a sequence of events is an important public health problem. However, there is no commonly accepted definition of "regular." In our motivating example, a study of HIV self-testing behavior among men who have sex with men, the primary interest lies in determining the effect of an intervention on regularity of self-testing events. However, we only observe the presence or absence of any testing events in a sequence of predefined time intervals, not exact event times. Our goal is to develop methods for estimating parameters associated with a gamma renewal process when only binary summaries of the process are observed and baseline covariates are available. We propose two approaches to estimation and inference: first, an efficient likelihood-based method in the case where only the data on the first interval containing at least one testing event is utilized; and second, a quasi-likelihood approach that uses all of the available data on each subject. We conduct simulation studies to evaluate performance of the proposed methods and apply them to our motivating example. A discussion on interesting directions for further research is provided.

EMAIL: john_rice@urmc.rochester.edu

10k. One-Step and Two-Step Estimation in a Time-Varying Parametric Model

Mohammed R. Chowdhury*, Kennesaw State University
Bradley Barney, Brigham Young University

Because the popularity of dynamic models continues to grow, and because nonparametric smoothing plays a preeminent role in the estimation of time-varying parameters, it is important to understand the properties of available estimation techniques. We consider the situation in which the response variable follows a parametric model indexed by a parameter that varies smoothly over time. Smooth estimates may be obtained via kernel smoothing or local polynomial estimation, among other possibilities. Furthermore, there are one-step implementations, which directly produce smoothed estimates, and two-step implementations, which first obtain raw parameter estimates on a grid of time values and then apply smoothing strategies directly to these raw quantities. We detail properties such as asymptotic biases, variances and mean squared errors of some such estimators. Application of one- and two-step smoothing procedures is demonstrated by large demographic studies. Additionally, we present a comparative simulation study that assesses one- and two-step smoothing estimation in terms of bias, MSE and smoothing estimates.

EMAIL: mchowd10@kennesaw.edu

10l. Estimating Complex Relationships with Nonlinear Longitudinal Growth Patterns

Brianna C. Heggeseth*, Williams College

One key benefit of a longitudinal study is the ability to observe and record how variables change over time. The individual pattern of growth and development may serve as a predictor of a future clinical measure or the result of early-life environmental or genetic factors. We compare existing tools using simulation studies as well as real childhood growth data and propose concrete modeling strategies within the generalized linear model and mixture model frameworks to estimate potentially complex relationships with nonlinear growth patterns.

EMAIL: bch2@williams.edu

10m. Bayesian Analysis of Multiple Longitudinal Outcomes of Mixed Types with Nonparametric Treatment Effects

Sheng Luo*, University of Texas Health Science Center at Houston
Jun Zhang, University of Texas Health Science Center at Houston

Impairment caused by Parkinson's disease (PD) affects multiple domains (e.g., motor, cognitive, and behavioral). Due to the heterogeneous nature and unknown pathogenic mechanisms of PD, clinical trials in PD collect multiple categorical and continuous longitudinal health outcomes to test the hypothesis that

the target treatments are more effective than placebo in slowing PD progression. In this project, we develop a latent trait linear mixed modeling framework based on dimension reduction and introduce a general definition of the treatment effects. We also develop a formal nonparametric model to define and estimate the time-dependent treatment effects. This method development has been motivated by and applied to the Neuroprotection Exploratory Trials in Parkinson's Disease (PD) Long-term Study-1 (LS-1 study, n = 1741), developed by The National Institute of Neurological Disorders and Stroke Exploratory Trials in Parkinson's Disease (NINDS NET-PD) network.

EMAIL: sheng.t.luo@uth.tmc.edu

11. POSTERS: Survival Methods

11a. A Decrease of Power for Progression-Free Survival in Patients with Early Progressive Cancer

Takanori Tanase*, Taiho Pharmaceutical Co., Ltd.
Chikuma Hamada, Tokyo University of Science

Progression-free survival is evaluated according to a tumor assessment schedule specified in the protocol. Oncology clinical trials aimed new drug developments are often conducted for later-line treatment patients with early progressive cancer. We focused on the fact that the hazard ratio for progression-free survival was significant; however, the median progression-free survival was almost the same between the treatment groups in clinical trials for patients with metastatic colorectal cancer refractory to standard chemotherapy. To quantitatively assess biases in hazard ratio and power for progression-free survival resulting from different tumor assessment schedules, we performed a computational simulation under the assumption of biomarker having interaction effect to treatment. The simulation results showed that there were biases in hazard ratio and the power was decreased in cases that median progression-free survival was short, tumor assessment schedule was largely spaced, and interaction effect of biomarker. Hazard ratio and power for progression-free survival is affected by tumor assessment schedule and biomarker in assumption of early progressive cancer.

EMAIL: t-tanase@taiho.co.jp

11b. Outcome-Dependent Sampling 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

Epidemiologic studies often seek to relate an exposure variable to a failure time that suffers from interval-censoring. When the failure is rare and time intervals are wide, a large cohort is usually required so as to yield reliable precision on the exposure-failure-time relationship. However, large cohort studies could be prohibitive for investigators with a limited budget, especially when the exposure variable is expensive to obtain. We propose a cost-effective outcome-dependent sampling design with interval-censored failure time data, where we enrich the observed sample by selectively including certain more informative failure subjects like the case-cohort design. We develop a novel sieve semiparametric maximum empirical likelihood approach to analyze data from the interval-censoring ODS design. This approach employs the empirical likelihood and sieve methods to handle the infinite-dimensional nuisance parameters, which greatly reduces the dimensionality and eases the computation difficulty. The consistency and asymptotic normality of the resulting estimator are established. Simulation results show that the proposed design and method work well.

EMAIL: qz4z3@mail.missouri.edu

11c. A Joint Model for Recurrent Events and a Semi-Competing Risk in the Presence of Multi-Level Clustering

Tae Hyun Jung*, Yale University
Heather Allore, Yale University
Denise Esserman, Yale University
Peter Peduzzi, Yale University

As clinical trial designs become complex, analytic techniques are necessary to address these complexities. For example, cluster designs can have multi-levels of clustering in which patients are nested within clinical sites (Level 2) and outcomes, such as recurrent events (e.g., falls), are nested within patients (Level-1). Patients can also experience

a semi-competing risk (e.g., death) which precludes the outcome of interest from being observed. Classic survival methods that analyze these processes separately may lead to erroneous inferences, as these two processes can be dependent under a multi-level structure. We developed a joint model that accounts for the association between recurrent events and a semi-competing risk in the presence of clustering at Level-1 and Level-2. Gaussian quadrature with a piecewise constant baseline hazard estimated the unspecified baseline hazards and the likelihood. Simulations show that the proposed joint model performs better (i.e., under 5% bias and 95% coverage) than shared frailty and joint frailty models when informative censoring and multi-levels of clustering exist. The proposed method is demonstrated using an AIDS clinical trial.

EMAIL: taehyun.jung@yale.edu

11d. A Modified Kolmogorov-Smirnov (MKS) Test for Detecting Differences in Survival Curves where there is a Delayed Treatment Effect

Rui Duan*, University of Pennsylvania
Jiaxin Fan, University of Pennsylvania
Meghan Buckley, University of Pennsylvania
Phyllis A. Gimotty, University of Pennsylvania

In clinical studies investigating immunotherapies, treated patients appear to have a delayed response to treatment. In this case, the Kaplan-Meier survival curves do not diverge until the time when patients respond to the treatment. Commonly used statistical tests such as the log-rank and Wilcoxon tests may be less powerful. Existing publications have limited discussion of this case and some methods require specification of the time to response or making extra assumptions. We investigate the performance of the MKS test (Fleming et al. 1980) to detect a difference in this situation. We use a data-generating process to simulate survival data with a delayed treatment response using a random response time for each subject. The performance of the MKS test is investigated using a simulation study and it is compared to the performance of the log-rank and Wilcoxon tests. The power of these tests under various alternative hypotheses is considered. By varying the time to response, the maximum difference between the survival curves and the underlying distribution of survival time, we identify the scenarios where the MKS test is more powerful than the log-rank and Wilcoxon tests.

EMAIL: ruiduan@mail.med.upenn.edu

11e. Joint Modeling of Event Times with Length-Bias and Competing Risks: Application to First and Second Stages of Labor in Pregnant Women

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

The progression of labor in pregnant women has long been a challenge for obstetricians, especially the first stage of labor during which the woman dilates from 0 cm to 10 cm and the second stage of labor during which the woman delivers the baby. Assessing the distributions of the durations of the two stages as well as their association is of considerable interest. Given that women are only observed after they get admitted to hospital when they are already dilated to a certain level, the data we have for the first stage duration is length-biased. Also, the second stage of labor could be terminated due to delivery of the baby or some possible complications. Thus, the duration of delivering stage has competing risks. We propose a joint model for the durations of the two stages which takes into account the unknown truncation time of the first stage duration as well as the competing risks in the second stage. The proposed approach also allows for prediction of second stage duration based on the first stage characteristics and a fitted joint model. We illustrate our proposals on a longitudinal labor data.

EMAIL: mlbegood@gmail.com

11f. Nonparametric Group Sequential Methods for Evaluating Survival Benefit from Multiple Short-Term Follow-Up Windows

Meng Xia*, University of Michigan
Susan Murray, University of Michigan
Nabihah Tayob, University of Texas MD Anderson Cancer Center

In this research, we take a fresh look at group sequential methods applied to two-sample tests of censored survival

data and propose an alternative method of defining and evaluating treatment benefit in this context. Following insight from health economists that patients favor short term over long term outcomes, we re-purpose traditional censored survival data collected over long follow-up periods into a sequence of (potentially overlapping) short-term follow-up windows, each holding information on short-term outcomes of interest. This restructuring of follow-up data, and application of corresponding two-sample restricted means tests described by Tayob and Murray, gives an alternative understanding of treatment benefit with efficiency gains. As part of developing group sequential methods for these analyses, we describe asymmetric error spending approaches that differentially limit the chances of stopping incorrectly for perceived efficacy versus perceived harm attributed to the investigational arm. Recommendations for how to choose proper group sequential stopping boundaries in different research settings are given, with supporting simulations and an example.

EMAIL: summerx@umich.edu

11g. Shape Restricted Additive Hazard Models

Yunro Chung*, Fred Hutchinson Cancer Research Center
Anastasia Ivanova, University of North Carolina, Chapel Hill
Michael G. Hudgens, University of North Carolina, Chapel Hill
Jason P. Fine, University of North Carolina, Chapel Hill

We consider estimation of the semiparametric additive hazards model with a unspecified baseline hazard function where the effect of a continuous covariate has a specific form of a shape but otherwise unspecified. It is particularly useful for a unimodal hazard function, where the hazard is monotone increasing and monotone decreasing along with a mode. The mode is generally unknown and needed to be estimated. A popular approach of the proportional hazards model may be limited in such setting owing to the complicated structure of the partial likelihood. Alternatively, our model defines a quadratic loss function, and its simple structure allows a global Hessian matrix that does not involve parameters. Thus, once the global Hessian matrix is computed, a standard quadratic programming method can be performed by profiling all hypothetical locations of the mode. On the other hand, the dimension of the global Hessian matrix is large, and we alternatively propose the quadratic pool-adjacent-violators algorithm to reduce computational costs. Analysis of data from a recent cardiovascular study illustrates the practical utility of our methodology.

EMAIL: yunro.roy@gmail.com

11h. A Hierarchical Generalized Gamma Model for Heavy-Tailed Survival Data: Large-Scale Application to Reporting Delays from Real-Time Disease Surveillance

Krzysztof Sakrejda, University of Massachusetts, Amherst
Stephen A. Lauer, University of Massachusetts, Amherst
Justin Lessler, Johns Hopkins Bloomberg School of Public Health
Nicholas G. Reich*, University of Massachusetts, Amherst

Surveillance systems generate data on disease incidence at a delay but public health decisions require current data. With appropriate adjustments for reporting delays, under-reported surveillance data can be scaled up in real-time to estimate current levels of a particular disease. In national digital surveillance systems, analysts face several challenges in estimating reporting delays, namely heavy-tailed reporting delay distributions and hierarchical administrative structures. We propose a framework for fitting a three-parameter generalized gamma model that captures the heavy-tails of reporting delays and the nested structure of the data. This model provides estimates of reporting delays, identifying low- and high-performing locations. Additionally, estimated reporting delays can be applied to real-time, under-reported data to estimate the current disease burden. We demonstrate the large-scale application of these methods using data from the Thai notifiable disease surveillance system for the years 2013-2015. We compare parameter estimates and real-time predictions at different spatial scales based on the model described above and a model that uses empirical quantiles.

EMAIL: nick@schoolph.umass.edu

11i. Cox Regression Model with Doubly Truncated Data

Lior Rennert*, University of Pennsylvania
Sharon X. Xie, University of Pennsylvania

Truncation is a well-known phenomenon that may be present in observational studies of time-to-event data. While methods exist for applying the Cox regression model to either left or right truncated data, no method exists for adjusting the model for simultaneous left and right truncation. We propose a weighted Cox regression model to adjust for double truncation, where the weights are estimated both parametrically and nonparametrically and are inversely proportional to the probability that a subject is observed. The resulting weighted estimator of the hazard ratio is consistent, asymptotically normal, and a consistent estimator of the asymptotic variance is provided when weights are estimated parametrically. We apply the bootstrap technique to estimate the variance when the weights are estimated nonparametrically. We demonstrate through extensive simulations that the proposed estimator has little bias, while the estimator resulting from the unweighted Cox regression model which ignores truncation is biased. We illustrate our approach in an analysis of autopsy-confirmed Alzheimer's disease patients to assess the effect of education on survival and time to symptom onset.

EMAIL: lior.rennert@gmail.com

11j. Covariate Dependent Cross-Ratio Analysis of Bivariate Time-to-Event Data

Ran Liao*, Indiana University
Tianle October Hu, Eli Lilly and Company

Cross-ratio, formulated as the ratio of two conditional hazard functions, offer a direct measure of dependence between two survival times that can account for censoring and accommodate potential covariates. Inherited the nice interpretation of hazard ratio from survival analysis setup, cross-ratio can be interpreted as the hazard ratio of one event conditional the status of the other event.

EMAIL: ranliao@iu.edu

11k. A Semiparametric Regression Model for Joint Cumulative Incidence Function

Xingyuan Li*, University of Pittsburgh
(Joyce) Chung-Chou H. Chang, University of Pittsburgh

In clinical studies, researchers are often interested in estimating the association between two correlated event times where each of these event times is subject to competing risks. Cheng et al. (2007) defined a joint cumulative incidence function (CIF) for bivariate failure times and proposed a nonparametric method to estimate the joint CIF. Under the proportional subdistributional hazards assumption, we propose a regression model to estimate the joint CIF allowing each of the two marginal subdistributional hazards functions to be dependent on covariates. Our proposed model extends the method proposed by Dixon et al. (2011) to allow for different failure types.

EMAIL: xil143@upitt.edu

11l. Test for Stratified Random Signs Censoring in Competing Risks

Shannon M. Woolley*, University of Pittsburgh
Jonathan G. Yabes, University of Pittsburgh
Abdus S. Wahed, University of Pittsburgh

In the setting of competing risks, the marginal survival functions of latent failure times are nonidentifiable without making further assumptions, the majority of which are untestable. One exception is random signs censoring which assumes the main event time is independent of the indicator that the main event preceded the competing event. Few methods exist to test this assumption, and none consider a stratified test, which detects whether random signs censoring is met within subgroups of a covariate. We develop a nonparametric stratified test for random signs censoring that is asymptotically normal and computationally simple. Through Monte Carlo simulations, we show our proposed test statistic has empirical levels close to the nominal level and maintains fairly high power with censoring. Compared to the unstratified test, our test has nearly equivalent power under random signs censoring and is superior in situations of stratified random signs censoring, where the unstratified test fails to detect random

signs censoring within subgroups. Its ease of implementation and utility are illustrated through an application to transplant data from the United Network for Organ Sharing.

EMAIL: smw81@pitt.edu

11m. Power Calculations for Interval-Censored Survival Analysis using Multiple Imputation Approach

Monica Chaudhari*, University of North Carolina, Chapel Hill
Edwin H. Kim, University of North Carolina, Chapel Hill
Michael R. Kosorok, University of North Carolina, Chapel Hill

Interval-censored survival data encountered in periodic assessments during a clinical trial involve measurements based on the time window within which a response variable crosses a fixed threshold. In contrast to this setting, we study the time to achieve a targeted allergenic food tolerance based on separately measured initial and subsequent thresholds that are both interval-censored. The Double-Blind Placebo-Control Food Challenge test is the gold standard in diagnosis of food allergy that exposes a patient to a fixed sequence of increasing dose levels of an allergen until a clinical symptom is elicited. This obscures a patient's true threshold between two consecutively administered doses, the highest with no symptom and the lowest with first. The Peanut SLIT-TLC clinical trial offers an opportunity to study the time to tolerance to a targeted dose by conducting two consecutive DBPCFCs. By virtue of the study design, the observed time is also interval-censored and independent of the initial threshold and subject characteristics. We propose multiple imputation methods for Cox regression of the threshold-crossing time and study their power with differing parameters.

EMAIL: monica.chaudhari@gmail.com

12. POSTERS: Cancer Applications

12a. Utility-Based Designs for Randomized Comparative Trials with Categorical Outcomes

Thomas A. Murray*, University of Texas MD Anderson Cancer Center
Peter F. Thall, University of Texas MD Anderson Cancer Center
Ying Yuan, University of Texas MD Anderson Cancer Center

General utility-based testing methodology for design and conduct of randomized comparative clinical trials with categorical outcomes is presented. Numerical utilities of all elementary events are elicited to quantify their desirabilities. These numerical values are used to map the categorical outcome probability vector of each treatment to a mean utility, which is used as a one-dimensional criterion for comparative tests. Bayesian tests are presented, including fixed sample and group sequential procedures, assuming Dirichlet-multinomial models for the priors and likelihoods. Guidelines are provided for establishing priors, eliciting utilities, and specifying hypotheses. Efficient posterior computation is discussed, and algorithms are provided for jointly calibrating test cutoffs and sample size to control overall type I error and achieve specified power. Asymptotic approximations for the power curve are used to initialize the algorithms. The methodology is applied to re-design a completed trial that compared two chemotherapy regimens for chronic lymphocytic leukemia using a design that dichotomized an ordinal efficacy outcome and ignored toxicity.

EMAIL: tamurray@mdanderson.org

12b. Network Meta-Analysis for Longitudinal Data with Heterogeneous Reporting

Hsin-Hui Huang, University of Pittsburgh
Joyce Chang, University of Pittsburgh
Emma Barinas-Mitchell, University of Pittsburgh
Brenda Diergaarde, University of Pittsburgh
Marnie Bertolet*, University of Pittsburgh

Traditional meta-analyses combine studies to determine the effect of an intervention on an outcome measured once at the

study's end. Bayesian network meta-analysis (NMA) methods expand this to allow longitudinal repeated outcome measures and a network of target treatments, including studies with any subset of the target treatments. There still exist difficulties in the practical implementation of these NMA methods; specifically: 1) heterogeneous reporting, with some studies reporting outcome as the mean and variance and others reporting the mean and variance of the change from baseline, and 2) no consistency measures exist for longitudinal NMA. In this study, starting with a Bayesian NMA for repeated outcome measures, we incorporated a first-order autoregressive model to convert the change from baseline variances to variances of the means. We updated and incorporated a traditional arm-based consistency method to determine whether the direct and indirect effect comparisons of treatments on repeated-measure outcomes were similar. We illustrate our method comparing breast cancer hormone therapies on changes in total cholesterol measured multiple times during follow-up.

EMAIL: mhb12@pitt.edu

12c. Estimating the Probability of Clonality Relatedness in Cases with Two Tumors

Audrey Mauguen*, Memorial Sloan Kettering Cancer Center
Venkatraman E. Seshan, Memorial Sloan Kettering Cancer Center
Irina Ostrovnaya, Memorial Sloan Kettering Cancer Center
Colin B. Begg, Memorial Sloan Kettering Cancer Center

When two tumors arise in a patient, a key question is to determine whether they are independent primary tumors, or whether one is a metastasis from the other. This can be done by comparing the mutations characterizing each tumor. Previously a test was developed for the null hypothesis that the tumors are independent. We have reframed the problem in the context of diagnosis using a random effects model. Taking advantage of all available information, the model estimates the proportion of clonal cases and the distribution of individual clonality signals representing the proportion of clonal mutations among all mutations in the tumor pair. This quantity is 0 when the tumors are independent. Here, we assume a lognormal distribution for the random effects among clonal cases. Estimated parameters are used to estimate the individual probabilities of being clonal for each case. Performance of the estimation is assessed through simulations. Impact of model misspecification was assessed by simulating data using a Beta distribution while estimating the parameters using the proposed lognormal model. The model is applied to pairs of in situ and invasive tumors in breast cancer patients.

EMAIL: mauguena@mskcc.org

12d. Spatial Bayesian Survival Analysis of SEER Breast Cancer Data

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

Cox proportional hazards models are commonplace in survival analysis and offer many advantages in the way of methods and interpretation. However, the proportional hazards assumption is not always met, and for these situations, the Accelerated Failure Time model is a good alternative with its own set of benefits. In this study, we explore the advantages and disadvantages of utilizing these differing approaches when also considering spatial frailty terms to account for unmeasured environmental exposure in Louisiana SEER breast cancer data. Ultimately, we aim to combine a model of this type with existing methods to increase breast cancer risk prediction accuracy.

EMAIL: rachel.carroll@nih.gov

12e. The Role of BMI at Diagnosis on Black-White Disparities in Colorectal Cancer Survival: A Counterfactual Density Regression Approach

Katrina L. Devick*, Harvard University
Linda Valeri, Harvard University
Jarvis Chen, Harvard University
Brent Coull, Harvard University

To date, a counterfactual framework has not been used to study the effect of a distribution-level shift in a continuous intermediate variable when considering racial/ethnic inequal-

ities in cancer. In the case of colorectal cancer, disparities in survival and body mass index (BMI) among non-Hispanic Whites and non-Hispanic Blacks are well documented. We quantify the impact of the hypothetical intervention on the distribution of BMI for Black colorectal cancer patients to match the distribution of BMI for White colorectal cancer patients. We estimate the residual disparity in survival after this intervention in a sample of colorectal cancer patients enrolled in the Cancer Care Outcomes Research and Surveillance (CanCORS) consortium, via a nonparametric Bayesian density regression for BMI and a parametric accelerated failure time model for survival. A simulation study is performed to compare the efficiency and bias of the residual disparity estimates when implementing a distribution-level shift of the mediator, BMI, relative to a location only shift, or a shift in categories for different underlying distributions of BMI.

EMAIL: khartzler@fas.harvard.edu

12f. Design Considerations in Integrated Randomized Phase II/III Oncology Clinical Trials

WITHDRAWN

Chen Hu*, Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins University School of Medicine

Randomized phase II trials have been increasingly used recently in oncology research. Integrated randomized phase II/III trial designs, which combines the standard randomized phase II and III aspects into a single study by implementing an adaptive design, thus have been increasingly considered as a potential alternative to accelerate clinical development. In this paper, we review the differences and investigate the performance of different design strategies of conducting phase II and III trials seamlessly or not. We particularly examine the impacts of association between endpoints and the choice of phase II endpoints on design performance in terms of type 1 error, power, study duration and expected sample size. We demonstrate that under appropriate conditions, in comparison of standard paradigm, integrated phase II/III design may achieve substantial gains in terms of sample size reduction, time and resource savings and shorter study durations.

EMAIL: huc@jhu.edu

12g. Innovative Phase I Dose-Finding Designs for Molecularly Targeted Agents and Cancer Immunotherapies

Jun Yin*, Mayo Clinic
Daniel J. Sargent, Mayo Clinic
Sumithra J. Mandrekar, Mayo Clinic

Phase I trials traditionally aim at determining the maximum tolerated dose (MTD) using grade 3+ toxicity data from cycle 1 only. These designs originate from anticancer therapies involving cytotoxic agents and may not be relevant for molecularly targeted agents (MTAs) and immunotherapies where moderate and late toxicities may become challenging. We recently have developed innovative designs to potentially overcome these challenges. In this work, we will review these new strategies, specifically, in three aspects: (1) development of novel toxicity endpoints summarizing multiple toxicity grades and types; (2) extension to repeated measurement of toxicity over multiple treatment cycles with time trend detection; and (3) incorporation of early efficacy signals in dose determination. We anticipate our approaches will gain greater acceptance and integration in practice, as well as motivating more alternative statistical proposals to utilize the complex data generated by phase I trials.

EMAIL: vivien.jyin@gmail.com

12h. A Joint Model of Cancer Incidence and Screening

Sheng Qiu*, University of Michigan
Alex Tsodikov, University of Michigan

Introduction of screening for prostate cancer using the prostate-specific antigen (PSA) biomarker of the disease in the late 80ies led to remarkable dynamics of the incidence of the disease. To explain the dynamics, we formulate a joint model of cancer progression to symptomatic (clinical) diagnosis and the screening process and the associated detection modality, as both processes interact to produce the observed incidence. The risks of screening and clinical diagnosis are dependent sharing the latent tumor onset and progression processes in the subject. Intensity of screening and the hazard driving prostate cancer progression are estimated jointly and semiparametrically using the NPMLE method based on the joint model. Asymptotic and nite sample properties of the proposed estimators are studied

analytically and by simulations. An application using data from the European cancer registry EUREG is presented.

EMAIL: shqiu@umich.edu

12i. A Dependent Dirichlet Process Model for Competing Risks Data with an Application to a Breast Cancer Study

Yushu Shi*, Medical College of Wisconsin
Purushottam Laud, Medical College of Wisconsin
Joan Neuner, Medical College of Wisconsin

Bone fractures are among the most common morbidities affecting breast cancer survivors, and there is evidence to suggest that hormonal therapy drugs may contribute to worsening bone densities. The aim of this paper is to compare the effects on fracture of two hormone therapy drugs: Tamoxifen, and Aromatase Inhibitors (AIs) with death as a competing risk. In order to model the non-constant subdistribution hazard ratio, we propose a dependent Dirichlet process competing risks model based on the Dirichlet process mixture of Weibulls model and the Fine and Gray model. In addition, we incorporate external time dependent covariate into our model. The flexibility of Bayesian nonparametric model enables us to address important clinical questions that are not easily provided by traditional methods. This paper first describes the model and compares it with existing methods through simulations, and then applies it to the motivating breast cancer dataset.

EMAIL: shiyushu2006@gmail.com

12j. A Comparison of Statistical Methods for the Study of Etiologic Heterogeneity

Emily C. Zabor*, Memorial Sloan Kettering Cancer Center
Colin B. Begg, Memorial Sloan Kettering Cancer Center

Epidemiologic research has traditionally been guided by the premise that certain diseases share an underlying etiology, or cause. However, with the rise of molecular and genomic profiling attention has increasingly focused on identifying subtypes of disease. As subtypes are identified, it is natural to ask the question of whether they share a common etiology or in fact arise from distinct sets of risk factors. The concept of differing risk factors across subtypes of disease is known as etiologic heterogeneity. The epidemiologic questions of interest in this line of research include 1) whether a risk factor has the same effect across all subtypes and 2) whether risk factor effects differ across levels of each individual tumor marker of which the subtypes are comprised. A number of statistical models have been proposed to address these questions. To illuminate the similarities and differences among the proposed methods, and to identify any advantages or disadvantages, we employ a data application to elucidate the interpretation of model parameters and available hypothesis tests, and we perform a simulation study to assess bias in effect size, type I error, and power.

EMAIL: zabore@mskcc.org

12k. Modeling Longitudinal Count Data Using the Generalized Monotone Incremental Forward Stagewise Method

Rebecca R. Lehman*, Virginia Commonwealth University
Colleen Jackson-Cook, Virginia Commonwealth University
Kellie J. Archer, The Ohio State University

When considering high-throughput genomic data, there is a growing need for predictive modeling methods when the number of explanatory variables exceeds the number of samples. Currently, there are few statistical methods for modeling a longitudinal count outcome in a high-dimensional predictor space. Often the count outcome may be a rate in which case Poisson or negative binomial(NB) regression are appropriate. We will present our extension of the Generalized Monotone Incremental Forward Stagewise Method to the longitudinal Poisson and NB regression models. Our methods will be applied to a breast cancer dataset in which Nuclear bud(NBud) count, a measure of chromosomal instability, was collected at 5 time points before and during cancer treatment. NBud frequency is determined by counting the number of binucleated cells with at least one NBud present in approximately 2,000 binucleated cells. Since previous research has implied that NBuds play a role in carcinogenesis, to elucidate molecular mechanisms involved in DNA damage, we are interested in identifying genomic features associated

with Nbuds. Thus, we will predict Nbud count using CpG site methylation and demographic data.

EMAIL: lehmanrr@vcu.edu

13. POSTERS: Machine Learning Methods

13a. Application of Advanced Data Mining Models to Identify Dietary Patterns Associated with Risk of Disease

Tinashe Michael Tapera*, Drexel University
Fengqing Zoe Zhang, Drexel University

Diet plays an important role in chronic disease. The use of pattern analysis in the overall diet is growing rapidly as a way of deconstructing the complexity of dietary intake and its relation to health. Pattern analysis methods have been applied to study the multiple dimensions of diet (e.g., PCA), however, existing methods do not fully utilize the predictive potential of assessment data and are sub-optimal at predicting clinically important variables. Thus, we propose examining dietary patterns using the advanced LASSO model to predict cardiovascular disease risk factors. Using Food Frequency Questionnaire data from the NHANES 2005-2006, we applied PCA and LASSO to identify dietary patterns in healthy US adults (n=2479) after controlling for confounding variables (e.g., age, BMI). Both analyses accounted for the sampling weights. Prediction was evaluated using cross-validation. Results found that the LASSO better predicts levels of triglyceride, LDL cholesterol, and total cholesterol (adjusted R squared = 0.889, 0.958, 0.986 respectively) than linear regression performed on the first 10 principle components (adjusted R squared = 0.021, 0.429, 0.645 respectively).

EMAIL: teetaps111@gmail.com

13b. A Semiparametric Method for Two-Group Classification

Seungchul Baek*, University of South Carolina
Yanyuan Ma, The Pennsylvania State University

In the classical discriminant analysis, one assumes two multivariate normal distributions with equal variance-covariance matrices. Under these conditions, the linear discriminant function is optimal with respect to maximizing the standardized difference between the means of two groups. However, for a typical case-control study, the assumption on distribution for case group needs to be relaxed in practice. Komori et al. (2015) proposed the generalized t-statistic to obtain a linear discriminant function with allowing for heterogeneity of case group. We propose a semiparametric method for two-group classification. This approach imposes no parametric assumptions on the probability distribution of case group even though normality assumption on control group would be reasonable. When linearity and constant variance assumption are satisfied, we showed our semiparametric estimator is efficient, compared to one proposed by Komori et al. (2015). We conduct simulation studies and a real data example is illustrated to show the finite sample performance.

EMAIL: sbaek@email.sc.edu

13c. TSPTree: A Nonparametric Machine Learning Classifier for Robust Cross-Platform Prediction in Genomic Applications

Chien-Wei Lin*, University of Pittsburgh
George Tseng, University of Pittsburgh

Classification algorithm is a widely-used machine learning technique to predict disease diagnosis in biomedical research. In addition to the disease prognosis, to understand the underlying pathological mechanism is also important. A robust and interpretable classifier with high reproducibility is always favored. The top scoring pair (TSP) algorithms is one example of simple rank-based algorithms to identify rank-altered gene pairs for classifier construction. Tree algorithms, for examples, CART, random forest, are also widely used in biomedical research and offer tree-like hierarchical model interpretation compared to other types of classification algorithms. Gene pathway database (e.g., KEGG, GO) provide knowledge-based insights to interpret the complex biological system across multiple genes. We propose a nonparametric classification model by combining TSP and tree-based classification algorithms, and incorporating existing pathway databases to guide the selections of gene pairs. We use simulations and real applications to lung cancer data to

demonstrate its robust and accurate prediction accuracy when applied to new cohorts using different platforms.

EMAIL: masaki396@gmail.com

13d. Compound Latent Dirichlet Allocation

Clint Pazhayidam George, University of Florida
Wei Xia*, University of Florida
George Michailidis, University of Florida

Topic models such as Latent Dirichlet allocation (LDA) are often used to make inference regarding the underlying topic structure of a corpus (e.g. a document collection). We imagine that a corpus is partitioned into several collections. Each of these collections shares a common set of topics, but there exists relative variation in the proportion of topics among collections. To incorporate any prior knowledge about these data organization hierarchy and metadata, we propose the compound latent Dirichlet allocation (cLDA) model. The parameters of interest in the model are hidden and inferred via posterior inference. As exact posterior inference is intractable for cLDA, we develop three approximate inference schemes based on the principles of Markov chain Monte Carlo, Langevin dynamics and variational inference. We evaluate the model performance using real-world datasets.

EMAIL: jjtwwdei@ufl.edu

13e. Estimating an Inverse Mean Subspace

Jiaying Weng*, University of Kentucky
Xiangrong Yin, University of Kentucky

We develop a new method called Fourier transformation inverse regression in the sufficient dimension reduction to capture the directions in the central subspace. Under linearity condition, the space spanned by Fourier transformation inverse regression would be in the central subspace. By forming a kernel matrix using Fourier transformation. we construct dimensional test via weighted chi-squared distribution. Our new method performs better than existing methods in the simulation study and real data analysis.

EMAIL: jiaying.weng@uky.edu

13f. Clustering Analysis on Identification of Diverse Sepsis Phenotypes

Zhongying Xu*, University of Pittsburgh
Hernando Gomez, University of Pittsburgh
Chung-Chou H. Chang, University of Pittsburgh

Sepsis is a life-threatening syndrome caused by the body's overwhelming inflammatory response to infection. Traditionally, sepsis has been considered as one syndrome with clinical presentations varying only by severity. However, recent data has challenged this paradigm, and has suggested that sepsis probably encompasses multiple phenotypes. In this study, we focused on exploring diverse patterns of sepsis in a cohort of critically ill patients, in order to better understand different response and further design targeted treatment on specific phenotypes. The dataset we used includes clinical, laboratory, and demographic information on adult patients admitted to a tertiary care institution with 8 intensive care units (ICU) in 8 years. Several clustering methods were applied including k-means based consensus clustering, tight clustering, and hierarchical clustering. Comparing different clustering methods, we found the number of clusters and the patterns describing these clusters are very similar. In addition, we also found that the resulting clusters were associated with subsequent clinical characteristics showing the clinical value of the identified clusters.

EMAIL: zhx17@pitt.edu

13g. Comparison of Linear and Non-Linear Models for Predicting Energy Expenditure from Raw Accelerometer Data

Alexander H.K. Montoye, Alma College
Munni Begum*, Ball State University
Zachary Henning, Ball State University
Karin A. Pfeiffer, Michigan State University

This study had three purposes, all related to evaluating energy expenditure (EE) prediction accuracy from body-worn accelerometers: 1) compare linear regression to linear mixed models, 2) compare linear models to artificial neural network models, and 3) compare accuracy of accelerometers placed on the hip, thigh, and wrists. Forty individuals performed 13 activities in a 90-minute semi-structured, laboratory-based

protocol. Participants wore accelerometers on the right hip, right thigh, and both wrists and a portable metabolic analyzer (EE criterion). For studies using wrist-worn accelerometers, machine learning models offer a significant improvement in EE prediction accuracy over linear models. Conversely, linear models showed similar EE prediction accuracy to machine learning models for hip- and thigh-worn accelerometers and may be viable alternative modeling techniques for EE prediction for hip- or thigh-worn accelerometers.

EMAIL: mbegum@bsu.edu

13h. Non-Parametric Cluster Significance Testing with Reference to a Unimodal Null Distribution

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

Cluster analysis is an unsupervised learning strategy that can be employed to identify subgroups of observations in data sets of unknown structure. This strategy is particularly useful for analyzing high-dimensional data such as microarray gene expression data. Many clustering methods are available, but it is challenging to determine if the identified clusters represent truly distinct subgroups rather than noise. We propose a novel strategy to investigate the significance of identified clusters by comparing the within-cluster sum of squares from the original data to that produced by clustering an appropriate unimodal null distribution. The null distribution we present for this problem uses kernel density estimation and thus does not require that the data follow any particular distribution. We find that our method can accurately test for the presence of clustering even when the number of features is high.

EMAIL: helgeson@live.unc.edu

13i. A Novel Deep Learning Classifiers using Brain Connectomic Circuitry Biomarkers as Input

Xiaoxiao Lu*, University of Maryland, College Park
Shuo Chen, University of Maryland, College Park

The differentially expressed brain connectivity circuitry biomarkers could play a critical role in mental disorder classification/prediction and treatment selection. However, different from the conventional individual biomarkers (e.g. individual genes) the network biomarker is an 'object' comprising correlated differentially expressed edges within a constrained spatial structure. In addition, the network biomarkers from different platforms are interacted with each other. To account for the hierarchical (edge, circuitry, multi-circuitry) and correlated input structure, we plan to develop a novel deep learning classifier by using combined state of the art 'deep-learning' and statistical techniques. We expect the new deep learning classifier to achieve improved prediction accuracy and robust performance, as the organized network biomarkers are more informative than conventional individual biomarkers.

EMAIL: xxluumd@gmail.com

14. POSTERS: Biomarkers

14a. Statistical Methodology for Comparing Biomarker-Guided Treatment Strategies

Meilin Huang*, University of Texas Health Science Center at Houston
Brian Paul Hobbs, University of Texas MD Anderson Cancer Center

Precision medicine has emerged from the awareness that human diseases are intrinsically heterogeneous with respect to their pathogenesis and composition among patient populations. Its application depends on our knowledge of distinct molecular profiles and identification of predictive biomarkers which can facilitate devising treatment (TX) strategies that exploit our current understanding of the mechanisms of the disease. We present a novel method for comparing the effectiveness of biomarker-guided strategies when evaluated with biomarker validation designs. Our method establishes a Bayesian framework for comparing TX strategies that integrate TX response surfaces, biomarker-guided allocation rules, and balance of prognostic characteristics between study cohorts. We also demonstrate how to correct cohort bias when randomization to cohort is infeasible. In simulation, comparing to five models which characterize the effects of predictive

biomarkers by TX-biomarker interactions using GLM and which handle inter-cohort effects using propensity score methods, our method reduced type I error in the prognostic setting, and increased power in the predictive setting.

EMAIL: meilin.huang@uth.tmc.edu

14b. Screening Approach for Testing SNP-SNP Interactions for a Binary Outcome

Huiyi Lin*, Louisiana State University Health Sciences Center
Po-Yu Huang, Industrial Technology Research Institute, Taiwan

We previously proposed the SNP Interaction Pattern Identifier (SIPI) approach, which can evaluate 45 various interaction patterns for a binary outcome using logistic regressions. The primary strengths of SIPI are taking non-hierarchical models and inheritance modes into consideration. Although the power of SIPI is better than several existing statistical approaches: the full additive model, SNPassoc and MDR-LR. The SIPI suffers from longer computation time, so an effective and efficient screening approach is needed for large scale studies. We propose SIPIscreen, a mini-version SIPI with the five full interaction models based on linear regression as the screening method for detecting SNP-SNP interaction. It has been shown that the significant tests in linear regression are similar compared with those in the logistic regression but faster. This is due to LPM can be estimated using ordinary least square instead of interactive estimation. This study conducted a simulation study to compare performance (type I error and power) and computation time of SIPIscreen and SIPI for detecting SNP-SNP interaction for a binary outcome.

EMAIL: hlin1@lsuhsc.edu

14c. Dynamic Prediction of Acute Graft-Versus-Host Disease (aGVHD) using Longitudinal Biomarkers

Yumeng Li*, University of Michigan
Thomas Braun, University of Michigan

Acute graft-versus-host disease (aGVHD) is a complication of allogeneic hematopoietic cell transplantation (aHCT) and is a leading cause of death in patients receiving aHCTs. Thus, investigators would like to have models that accurately predict those patients most likely to develop aGVHD in order to minimize over-treatment with steroids as well as reduce mortality. To this end, we propose using biomarkers that are collected weekly to predict the time-to-aGVHD through both joint modeling and landmark analysis. We consider settings in which the population is a mixture of various aGVHD risk classes so that the biomarker trajectories are irregular and possibly from different distributions. Unlike existing approaches, we start with no specification of the number of latent classes and explore the ideal number of risk classes suggested from the data. We compare different modeling approaches through simulations motivated from actual data collected at the University of Michigan Blood and Marrow Transplant Program.

EMAIL: yumeng@umich.edu

14d. A Joint-Modeling Approach to Assess Longitudinal Measurements of Biomarkers and Survival Outcomes for Melanoma Patients

Dawen Sui*, University of Texas MD Anderson Cancer Center
Yuling Wang, University of Texas MD Anderson Cancer Center
Jeffrey E. Lee, University of Texas MD Anderson Cancer Center

Biomarkers play a prominent role in medical decision making. Most studies that evaluate biomarkers focus on a single measurement. Serial measurements of biomarkers may provide important information about disease progression. We propose a joint modeling approach to evaluate the predictive effect of variability in serial measurements of biomarkers on survival outcomes for patients with melanoma. A two-stage modeling approach was applied in which the first stage modeled longitudinal, repeated measurements with unequal time measures and different numbers of follow-up visits, while the second stage modeled the time-to-event data. The random effect connects the two models under maximum likelihood estimation. A variety of integrative methods and random covariance structures were examined. Different methods for

model diagnostics were used to detect the model goodness of fit. Joint modeling was applied using 6 biomarkers from 120+ melanoma patients treated at MDACC from 1999 to 2007. Preliminary analyses demonstrated that levels of the biomarkers CRP, S100B, and vitamin D significantly correlated with survival outcomes. Larger studies are needed to confirm these findings.

EMAIL: dawensui@mdanderson.org

14e. Using Hierarchical Group Testing to Estimate the Prevalence of Multiple Diseases

Md Shamim Warasi*, Radford University
Joshua M. Tebbs, University of South Carolina
Christopher S. McMahan, Clemson University
Christopher R. Bilder, University of Nebraska, Lincoln

With the primary goal of reducing costs, group testing (pool testing) is widely used to screen individuals for sexually transmitted diseases. A recent shift in the group testing research community has seen the development of group testing estimation and case identification procedures for multiple diseases at once. The goal of this article is to propose new estimation techniques for general hierarchical group testing algorithms for multiple infections, both from frequentist and Bayesian perspectives. A Bayesian approach allows one to relax the potentially untrustworthy assumption that assay accuracy probabilities are known in advance, and it also allows us to incorporate historical information on disease prevalence. We illustrate our estimation methods using two data applications.

EMAIL: msarker@radford.edu

15. Evaluating Diagnosis Tests and Risk Prediction Using Clinical Utility Measures

› Decision Curve Analysis: Where are we 10 Years on?

Andrew J. Vickers*, Memorial Sloan Kettering Cancer Center

Decision curve analysis (DCA) was proposed in 2006 as a method to evaluate prediction models, diagnostic tests and markers. The key concept is to include a rational, decision analytic weighting of the practical consequences of using the model, test or marker in clinical practice (cf. AUC where sensitivity and specificity are weighted equally). DCA is a plot of net benefit (NB) against threshold probability (p) where NB = true positives-false positives x p/(1-p). Threshold probability is the minimum probability at which a doctor or patient would opt for intervention, hence positive vs. negative test status is defined as predicted risk greater to or equal to p. Theoretical developments since 2006 include application to censored data, correction for overfit and proof that NB is a proper scoring rule. DCA is now widely used in empirical practice and recommended by several top journals. Determining whether a model, test or marker is of clinical benefit requires that we incorporate clinical consequences into our statistical evaluations. DCA is a robust method for doing so and expanded practical use is warranted.

EMAIL: vickersa@mskcc.org

› Evaluating Markers for Risk Prediction: Decision Analysis to the Rescue

Stuart G. Baker*, National Cancer Institute, National Institutes of Health

There is considerable controversy over the choice of metric for evaluating an additional marker for a risk prediction model. Typically, the debate centers on the odds ratio, the change in area under the Receiver Operating Characteristic curve (AUC), and the net reclassification improvement, with no clear rationale to prefer any of these over the other two. Decision analysis offers a way forward by considering the clinical consequences of risk prediction. Decision curves and relative utility curves provide a sensitivity analysis based on the risk threshold (the probability of the event corresponding to indifferent between treatment and no treatment). The test tradeoff (at a risk threshold) is the minimum number of tests for an additional marker that need to be traded for a true positive to yield a positive net benefit. A simple decision-analytic metric for evaluating an additional marker is the approximate range of test tradeoffs, which is a function of only the AUC and the probability of the event.

EMAIL: sb16i@nih.gov

A Framework for Evaluating Precision Prevention

Holly Janes*, Fred Hutchinson Cancer Research Center

Precision prevention, or targeting the provision of prophylactic interventions to subpopulations, is of increasing importance in settings with resource-intensive interventions and given finite public health resources. We describe a framework for developing and evaluating precision prevention policies. Risk-based policies that target subgroups at high risk of disease, and which are in common use in many disease areas, are considered and contrasted with other policy approaches such as those targeting subgroups most likely to benefit from intervention. Policies that rely on individual-level variables, as well as policies reliant on group-level variables, are considered. We illustrate the approaches with application to two examples in an HIV prevention context.

EMAIL: hjanes@fhcrc.org

Benefit-Risk Evaluation for Diagnostics: A Framework (BED-FRAME)

Scott R. Evans*, Harvard University
Thuy T. Tran, Harvard University

The medical community needs systematic and pragmatic approaches for evaluating the benefit-risk trade-offs of diagnostics that assist in medical decision making. Benefit-Risk Evaluation of Diagnostics: A Framework (BED-FRAME) is a strategy for pragmatic evaluation of diagnostics designed to supplement traditional approaches. BED-FRAME evaluates diagnostic yield and addresses 2 key issues: (1) that diagnostic yield depends on prevalence, and (2) that different diagnostic errors carry different clinical consequences. As such, evaluating and comparing diagnostics depends on prevalence and the relative importance of potential errors. BED-FRAME provides a tool for communicating the expected clinical impact of diagnostic application and the expected trade-offs of diagnostic alternatives. BED-FRAME is a useful fundamental supplement to the standard analysis of diagnostic studies that will aid in clinical decision making.

EMAIL: evans@sdac.harvard.edu

16. Analysis Of Time-To-Event Data Subject To Length-Bias Sampling or Truncation

Prevalent Cohort Studies: Length-Biased Sampling with Right Censoring

Masoud Asgharian*, McGill University

Logistic or other constraints often preclude the possibility of conducting incident cohort studies. A feasible alternative in such cases is to conduct a cross-section prevalent cohort study for which we recruit prevalent cases, that is, subjects who have already experienced the initiating event, say the onset of a disease. When the interest lies in estimating the lifespan between the initiating event and a terminating event, say death, such subjects may be followed prospectively until the terminating event or loss to follow-up, whichever happens first. It is well known that prevalent cases have, on average, longer lifespans. As such, they do not form a random sample from the target population. If the initiating events are generated from a stationary Poisson process, this bias is called length bias. I present the basics of nonparametric inference using length-biased right censored data. I will then discuss some recent progress and current challenges.

EMAIL: masoud@math.mcgill.ca

Permutation Tests for General Dependent Truncation

Rebecca A. Betensky*, Harvard School of Public Health
Sy Han A. Chiou, Harvard School of Public Health
Jing Qian, University of Massachusetts, Amherst

Quasi-independence is a common assumption that enables simple analysis of truncated survival data that are frequently encountered in biomedical science, astronomy, and social science. Current methods for testing for quasi-independent truncation are powerful for monotone alternatives, but not otherwise. Extending methods in detecting highly non-monotone and even non-functional dependencies, we develop nonparametric tests that are powerful against nonmonotone alternatives. We describe and validate the use of an unconditional permutation method, which enables fixed risk-set-size permutation inference. The size and power of the proposed

testing procedure are assessed in extensive simulation studies. An aging study in cognitive and functional decline is included to illustrate the usefulness of the method.

EMAIL: betensky@hsph.harvard.edu

› Cox Regression for Doubly Truncated Data

Micha Mandel*, The Hebrew University of Jerusalem
Jacobo de Una Alvarez, University of Vigo

David K. Simon, Beth Israel Deaconess Medical Center and Harvard Medical School

Rebecca A. Betensky, Harvard School of Public Health

Doubly truncated data arise when event times are observed only if they fall within subject-specific, possibly random, intervals. While non-parametric methods for survivor function estimation using doubly truncated data have been intensively studied, only a few methods for fitting regression models have been suggested, and only for a limited number of covariates. In this talk, I present a method to fit the Cox regression model to doubly truncated data with multiple discrete and continuous covariates, and describe how to implement it using existing software. The approach is used to study the association between candidate single nucleotide polymorphisms and age of onset of Parkinson's disease.

EMAIL: micha.mandel@mail.huji.ac.il

› Smoothing Methods to Estimate the Hazard Rate under Double Truncation

Carla Moreira*, University of Vigo

Smoothing methods to estimate a doubly truncated density function were introduced and investigated in Moreira and de Uña-Álvarez (2012), Moreira and Van Keilegom (2013) and Moreira et. al (2016). However, in survival analysis another characteristic curve of much interest is the hazard rate function. The estimation of this function can be performed by kernel methods in the same spirit of the kernel density estimators. As for the density, proper corrections to deal with the problem of random double truncation are needed. In this setting, we will explore the performance of new kernel estimators for the hazard rate, including the investigation of the crucial issue of bandwidth selection. To this end, plug-in methods, cross-validation methods, and bootstrap methods will be considered. The methods are first shown to work from a theoretical point of view. A simulation study will be performed to assess the finite sample behavior of these bandwidth selectors. The use of the various practical bandwidth selectors by means of applications to medical data are included for illustration purposes.

EMAIL: carlamgmm@gmail.com

17. Issues and Solutions for Single-Cell RNA-seq Data Analysis

› On the Widespread and Critical Impact of Systematic Bias and Batch Effects in Single-Cell RNA-seq Data

Stephanie Hicks, Dana-Farber Cancer Institute
Mingxiang Teng, Dana-Farber Cancer Institute
Rafael A. Irizarry*, Dana-Farber Cancer Institute

Single-cell RNA-Sequencing (scRNA-Seq) has become the most widely used high-throughput method for transcription profiling of individual cells. Systematic errors, including batch effects, have been widely reported as a major challenge in high-throughput technologies. Surprisingly, these issues have received minimal attention in published studies based on scRNA-Seq technology. We examined data from five published studies and found that systematic errors can explain a substantial percentage of observed cell-to-cell expression variability. Specifically, we found that the proportion of genes reported as expressed explains a substantial part of observed variability and that this quantity varies systematically across experimental batches. Furthermore, we found that the implemented experimental designs confounded outcomes of interest with batch effects, a design that can bring into question some of the conclusions of these studies. Finally, we propose a simple experimental design that can ameliorate the effect of theses systematic errors have on downstream results.

EMAIL: rafa@jimmy.harvard.edu

SCnorm: A Quantile-Regression Based Approach for Robust Normalization of Single-Cell RNA-seq Data

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

Normalization of RNA-sequencing data is essential for accurate downstream inference, and a number of robust methods exist for bulk RNA-seq experiments. The assumptions upon which bulk methods are based do not hold for single-cell RNA-seq data and, consequently, applying bulk normalization methods in the single-cell setting introduces artifacts that substantially bias downstream analyses. To address this, we introduce SCnorm for accurate and efficient normalization of scRNA-seq data. Applications to a number of data sets demonstrate that SCnorm provides for more accurate estimates of fold-change as well as increased power and precision for downstream analyses.

EMAIL: kendzior@biostat.wisc.edu

A Unified Statistical Framework for RNA Sequence Data from Individual Cells and Tissue

Lingxue Zhu, Carnegie Mellon University
Jing Lei, Carnegie Mellon University
Bernie Devlin, University of Pittsburgh
Kathryn Roeder*, Carnegie Mellon University

Compared to tissue-level RNA-seq data, single cell sequencing yields valuable insights about gene expression profiles for different cell types, which is potentially critical for understanding many complex human diseases. However, developing quantitative tools for such data remains challenging because of technical noise, especially the dropout events. A dropout happens when the RNA for a gene fails to be amplified prior to sequencing, producing a false zero in the observed data. We propose a unified statistical framework for both single cell and tissue RNA-seq data, formulated as a hierarchical model. Our framework borrows the strength from both data sources and models the dropouts in single cell data, leading to a more accurate estimation of cell type specific gene expression profile. In addition, our model provides inference on (i) the dropout entries in single cell data that need to be imputed for downstream analyses; and (ii) the mixing proportions of different cell types in tissue samples. Simulation results illustrate that our framework outperforms existing approaches both in correcting for dropouts in single cell data, as well as in deconvolving tissue samples.

EMAIL: kathryn.roeder@gmail.com

Expression Recovery in Single Cell RNA Sequencing

Mo R. Huang, University of Pennsylvania
Mingyao R. Li, University of Pennsylvania
Nancy R. Zhang*, University of Pennsylvania

In single cell RNA sequencing experiments, not all transcripts present in the cell are captured in the library, and not all molecules present in the library are sequenced. The efficiency, that is, the proportion of transcripts in the cell that are eventually represented by reads, can vary between 2-60%, and can be especially low in highly parallelized droplet-based technologies where the number of reads allocated for each cell can be very small. This leads to a severe case of not-at-random missing data, which hinders and confounds analysis, especially for low to moderately expressed genes. To address this issue, we introduce a noise reduction and missing-data imputation framework for single cell RNA sequencing, which allows for cell-specific efficiency parameters and borrows information across genes and cells to fill in the zeros in the expression matrix as well as improve the expression estimates derived from the low read counts. We demonstrate the accuracy of the imputation through the subsampling of real high quality scRNA-seq data sets, and illustrate how this critical imputation step improves downstream analyses in single cell experiments.

EMAIL: nzh@wharton.upenn.edu

› Global Prediction of Chromatin Accessibility Using RNA-Seq from Single Cell and Small Number of Cells

Weiqiang Zhou, Johns Hopkins Bloomberg School of Public Health
Zhicheng Ji, Johns Hopkins Bloomberg School of Public Health
Hongkai Ji*, Johns Hopkins Bloomberg School of Public Health

Conventional high-throughput technologies for mapping regulatory element activities such as ChIP-seq, DNase-seq and FAIRE-seq cannot analyze samples with small number of cells. The recently developed ATAC-seq allows regulome mapping in small-cell-number samples, but its signal in single cell or samples with ≤500 cells remains discrete or noisy. Compared to these technologies, measuring transcriptome by RNA-seq in single-cell and small-cell-number samples is more mature. Here we develop BIRD, Big Data Regression for predicting DNase I hypersensitivity using gene expression. We show that BIRD can globally predict chromatin accessibility and infer regulome using RNA-seq. Genome-wide chromatin accessibility predicted by RNA-seq from 30 cells is comparable with ATAC-seq from 500 cells. Predictions based on single-cell RNA-seq can more accurately reconstruct bulk chromatin accessibility than using single-cell ATAC-seq by pooling the same number of cells. Integrating ATAC-seq with predictions from RNA-seq increases power of both methods. Thus, transcriptome-based prediction can provide a new tool for decoding gene regulatory programs in small-cell-number samples.

EMAIL: hji@jhsph.edu

18. Comparative Effectiveness Research Methods Using EHR Data

› Mitigating Confounding Violations in Bayesian Causal Inference using High Dimensional Data

Jacob V. Spertus, Harvard Medical School
Sharon-Lise T. Normand*, Harvard Medical School

Increased access to health information has prompted ambitious attempts to understand the effect caused by new medical interventions in usual care populations. By conditioning on rich confounding information and utilizing larger populations, researchers aim to comply with key principles underpinning causal inference. We consider Bayesian regularization and model averaging techniques to effect estimation in high-dimensional data for binary treatments. We propose a two-step Bayesian propensity score weighting procedure using a beta-binomial based estimator of the difference in weighted pseudo-populations. Simulations indicate our two-step Bayesian propensity score weighting and Bayesian regularization methods are effective in estimating average treatment effects, improving on other options in terms of mean squared error. Bayesian propensity score methods also improved over frequentist weighting and doubly-robust estimation in high-dimensional settings. Registry data for about 9000 patients and nearly 500 confounders illustrate approaches by comparing drug eluting coronary stents to bare metal coronary stents on various outcomes. Funded by R01- GM111339 and U01-FDA004493.

EMAIL: sharon@hcp.med.harvard.edu

› Using Electronic Health Record Data for Basic Research and Discovery

William S. Bush*, Case Western Reserve University
Dana C. Crawford, Case Western Reserve University

The growing widespread adoption of electronic health record (EHR) systems enables new opportunities for biomedical research. Most notably, groups like the eMERGE network -- and the proposed National Precision Medicine Initiative – have and continue to amass large collections of individuals with detailed clinical information linked to biological specimens and their derived 'omics data. This amalgamation of biological and clinical data has enabled both new biomedical discoveries, such as the association of FOXE1 variants to hypothyroidism, and new study designs, such as Phenome-Wide Association Studies (PheWAS). In this talk, we will discuss the key benefits of using EHRs for basic research – access to large and unique study populations, detailed laboratory values, and rich clinical history. We will also discuss the drawbacks and perils of research driven by EHR data, including general clinical noise, inconsistency of measurement and follow-up, and incongruities between clin-

ical coding systems and biomedical diagnoses. Finally, we will review the state of PheWAS analyses and their findings.

EMAIL: wsb36@case.edu

› Old Bottle, New Wine, or New Bottle, Old Wine: Design Issues for Phenotyping in Electronic Health Record-Based Studies

Jinbo Chen*, University of Pennsylvania
Lu Wang, University of Pennsylvania
Rebecca Hubbard, University of Pennsylvania
Yong Chen, University of Pennsylvania

Systematic recording of rich health information in electronic formats has been widely adopted in the U.S. health system. This promises enormous flexibility and breath for clinical research at low cost and accelerated time frame through linkage with bio-repositories and other sources of genomic, biomarker, and environmental exposure data. The feasibility of clinical studies with electronic health record (EHR)-derived phenotypes has been demonstrated by successful replication of known genetic susceptibility variants for multiple complex phenotypes. But extraction of clinical phenotypes is not a trivial task. To study the association between a binary phenotype and a set of covariates, cases and controls are identified by applying selection rules, and therefore are subject to imperfect accuracy. A common practice is to validate the rule-based case-control status by performing medical chart review on a subset of study subjects. If necessary, this subset can subsequently be exploited to develop improved algorithms for case and control identification. We study efficient sampling methods for phenotype validation.

EMAIL: jinboche@mail.med.upenn.edu

› Using Historical, Electronic Health Record Data to Bolster a Cluster Randomized Trial of Early Childhood Obesity

Jonathan S. Schildcrout*, Vanderbilt University
Aihua Bian, Vanderbilt University

The Greenlight Study is a cluster randomized clinical trial that seeks to examine two strategies for reducing early childhood obesity in susceptible populations. Since the number of participating clinics was small, we retrospectively ascertained data from each EHR for non-participants during time periods that were prior to and conincident with Greenlight. By collecting historical EHR data, we have the opportunity to examine within clinic contrasts in treatment effectiveness. However, to improve the likelihood for eliminating participation bias, we rely on the data from non-participants who, in fact, received the clinic associated intervention while Greenlight was ongoing. In this talk, we will discuss longitudinal analyses of Greenlight and EHR data collected from non-participants. We will highlight several of the challenges and lessons learned from using EHR data to bolster this cluster randomized clinical trial of early childhood obesity.

EMAIL: jonathan.schildcrout@vanderbilt.edu

19. Health Policy Data Science

› Computational Health Economics for Identification of Unprofitable Health Care Enrollees

Sherri Rose*, Harvard University
Savannah L. Bergquist, Harvard University
Timothy Layton, Harvard University

We take the hypothetical role of a profit-maximizing insurer attempting to design its health plans to attract profitable enrollees and deter unprofitable ones. Such an insurer would not be acting in the interests of providing socially efficient levels of care by offering plans that maximize the overall benefit to society, but rather intentionally distorting plan benefits in order to avoid high-cost enrollees to the detriment of both health and efficiency. We focus on a specific component of health plan design: the prescription drug formulary, one of the most important dimensions on which insurers can distort their plan benefits in response to selection incentives as other dimensions are now highly regulated (e.g., pre-existing conditions). Our computational health economics approach centers around developing an ensembled machine learning method to determine which drug classes are most predictive of a new measure of unprofitability we derive, and thus most vulnerable to distortions by insurers in the Health Insurance Marketplaces. This work is designed to highlight vulnerable

unprofitable groups that may need special protection from regulators in health insurance market design.

EMAIL: rose@hcp.med.harvard.edu

› Precision Medicine: Statistical Methods to Improve Patient Outcomes and Support Value-Based Care

R. Yates Coley*, Group Health Research Institute

Precision promise to improve medical decision-making in the era of big data by making up-to-date analyses of patient information and scientific knowledge available to physicians and patients in real-time. In this talk, we will present a project from the Johns Hopkins Individualized Health Initiative that supports a personalized prostate cancer management program via a continuously learning prediction algorithm for the underlying cancer state. We propose a Bayesian hierarchical approach for extending local implementation of the algorithm to a multi-cohort setting and, in so doing, enabling estimation of population-level parameters. We discuss how such population-level parameter estimates from models motivated by prediction on the patient level can also be leveraged to inform health care policy.

EMAIL: coley.r@ghc.org

› Methods for Estimating the Health Effects of Air Pollution Sources in a Region

Roger Peng*, Johns Hopkins Bloomberg School of Public Health

Exposure to particulate matter (PM) air pollution has been associated with cardiovascular disease (CVD) hospitalizations and other clinical parameters. Information obtained from multisite studies is critical to understanding how PM impacts health and to informing local strategies for reducing individual-level PM exposure. However, few methods exist to perform multisite studies of PM sources and adverse health outcomes. We developed SHARE, a hierarchical modeling approach that facilitates reproducible, multisite epidemiologic studies of PM sources. SHARE is a two-stage approach that first summarizes information about PM sources across multiple sites. Then, this information is used to determine how community-level health effects of PM sources should be pooled to estimate regional-level health effects. Using data from 63 northeastern US counties, we estimated regional-level health effects associated with short-term exposure to major types of PM sources. We found PM from secondary sulfate, traffic, and metals sources was most associated with CVD hospitalizations.

EMAIL: rdpeng@gmail.com

› Transporting the Results of a Randomized Clinical Trial to a Target Population

Sarah Robertson*, Brown University
Issa Dahabreh, Brown University

We review methods for transporting the findings of a completed randomized controlled trial to a well-defined target population in which treatment and outcome information is unavailable. We discuss the evaluation of model assumptions, model selection, and the robustness of alternative methods to model misspecification. We present the results of simulation studies to assess the finite sample performance of different methods.

EMAIL: sarah_robertson@brown.edu

20. Statistical Advances in Scientific Reproducibility and Replicability

› Irreproducible Discovery Rate Regression

Feipeng Zhang, The Pennsylvania State University
Tao Yang, The Pennsylvania State University
Qunhua Li*, The Pennsylvania State University

When analyzing the reproducibility of findings from high-throughput experiments, one often treats all the candidates in one experiment in one global analysis. But this may be inappropriate in the scenarios that the reproducibility of candidates depends on some auxiliary variables. For example, in high-throughput chromosome conformation capture (Hi-C) data, a high-throughput technology for studying genome-wide chromatin structures, the reproducibility of interactions heavily depends on the proximity of the loci between which the interactions occur. A global analysis using existing reproducibility measures can lead to overly optimistic estimation of reproducibility and generate biased results. To address this issue, we introduce an approach called irrepro-

ducible discovery rate regression that incorporates the auxiliary information into the reproducibility assessment. By modeling the auxiliary variables through a regression framework, this method effectively improves the identification of reproducible signals and reduce the bias. We illustrate our method using both simulations and real data analyses on Hi-C data.

EMAIL: qunhua.li2@gmail.com

› A Framework for Discussing Reproduction and Replication of Scientific Studies

Prasad Patil*, Harvard School of Public Health
Roger D. Peng, Johns Hopkins Bloomberg School of Public Health
Jeffrey T. Leek, Johns Hopkins Bloomberg School of Public Health

There has been a declaration of crisis in science among the popular media. The inability to reproduce study results by rerunning published code on published data has raised eyebrows. Recently, large-scale efforts to replicate studies by collecting a second sample and reapplying a published protocol have yielded incendiary conclusions. Consortium studies from OSC and AMGEN have reported shockingly little concordance between multiple studies designed to examine the same scientific question. The specters of publication bias and outright fraud have cast a pall over psychological, social, and biological research. A large portion of the confusion surrounding reproducibility and replicability stems from a lack of a framework for defining and discussing these terms. We provide mathematical representations of the key aspects of a scientific study in an attempt to establish statistical definitions for reproducibility, replicability, and related concepts. We then illustrate recent replication efforts within our framework to provide further context, a standardized ontology for discussion, and to help determine if these results truly indicate a crisis in science.

EMAIL: prpatil42@gmail.com

› Inference Following Aggregate Level Hypothesis Testing in Large Scale Genomic Data

Ruth Heller*, Tel-Aviv University
Nilanjan Chatterjee, Johns Hopkins University
Abba Krieger, University of Pennsylvania
Jianxin Shi, National Cancer Institute, National Institutes of Health

In many genomic applications, it is common to perform tests using aggregate-level statistics within naturally defined classes for powerful identification of signals. Following aggregate-level testing, it is naturally of interest to infer on the individual units that are within classes that contain signal. Failing to account for class selection will produce biased inference. We develop multiple testing procedures that allow rejection of individual level null hypotheses while controlling for conditional (familywise or false discovery) error rates. We use simulation studies to illustrate validity and power of the proposed procedures in comparison to several possible alternatives. We illustrate the usefulness of our procedures in a natural application involving whole-genome expression quantitative trait loci (eQTL) analysis across 17 tissue types using data from The Cancer Genome Atlas (TCGA) Project.

EMAIL: ruheller@gmail.com

› Dynamic Statistical Comparisons and Extensible Research

Matthew Stephens*, University of Chicago
Gao Wang, University of Chicago

Empirical comparisons are a key way to assess the value of novel approaches. However, performing these kinds of comparisons is incredibly time-consuming, requiring careful familiarization with software and the creation of pipelines and scripts. In fast-moving fields new methods appear so frequently that comparisons are quickly out of date. To address this we are working on tools for creating "Dynamic Statistical Comparisons:" comparisons of methods with one another in a reproducible and easily-extensible way. The ultimate goal is public repositories that can provide "push-button" reproducibility of all comparisons. This talk will review our progress in

this direction, and other things we have learned about reproducible and extensible workflows along the way.

EMAIL: mstephens@uchicago.edu

21. Microbiome Research Methods

Compositional Mediation Analysis for Microbiome Studies

Michael Sohn*, University of Pennsylvania
Hongzhe Li, University of Pennsylvania

Motivated by advances in the causal inference on mediation and problems arising in the analysis of metagenomic data, we consider the effect of a treatment on an outcome transmitted through compositional mediators. Compositional and high dimensional features of such mediators make the standard mediation analysis not directly applicable. We propose a sparse mediation model for high-dimensional compositional data utilizing the algebraic structure of a composition under the simplex space and a constraint linear regression model to achieve subcompositional coherence. Under this model, we develop estimation method for estimating indirect microbial mediation effect and direct effect of a randomly assigned treatment on the outcome and variances using bootstrap. We conduct extensive simulation studies to assess the performance of our method and apply our method to a real metagenomic dataset to investigate the effect of fat intake on body mass index (BMI) transmitted through the gut microbiome.

EMAIL: msohn@mail.med.upenn.edu

Microbiome Normalization Methods: Effect on Dimension Reduction Analysis

Ekaterina Smirnova*, University of Montana
Glen Satten, Centers for Disease Control and Prevention
Snehalata Huzurbazar, University of Wyoming

Microbiome next generation sequencing experiments measure counts of DNA fragments for a large number of species in a sample. The total number of reads might vary dramatically across samples, and therefore appropriate scaling is necessary for any analysis. Prior to data analysis, normalization and bias adjustment are often implemented using filtering, library size adjustment and variance stabilization methods. Methods developed for RNA-seq data are being used for microbiome data, however the latter are extremely sparse and often dominated by a small number of species. Using methods from RNA-seq literature, and recently developed methods for adjustment of microbiome data, we illustrate the effects of normalization and bias adjustment on eigen-decomposition based dimension reduction analyses.

EMAIL: ekaterina.smirnova@mso.umt.edu

Penalized Logistic Regression Model for Microbiome Compositional Data

Jiarui Lu*, University of Pennsylvania
Hongzhe Li, University of Pennsylvania

We consider a penalized logistic regression model with a group of linear constraints to incorporate the compositional nature of the data and to achieve subcompositional coherence. Generalized Accelerated Proximal Gradient Method is used to estimate the regression coefficient and a de-biased procedure is proposed to obtain asymptotically unbiased and normal distributed de-biased estimates. Based on the asymptotic results, confidence intervals of the regression coefficients are obtained and used for variable selection. Simulations results show the correctness of the coverage probability of the confidence intervals and smaller variances of the estimates when using the groups of linear constraints. The true positive rates of selecting variables based on the confidence intervals are also studied to show the validity of the confidence interval-based variable selection. The proposed methods are applied to a gut microbiome dataset to identify bacterial genera that are associated with obesity and overweight and another data set to investigate the association between gut microbiome and inflammatory bowel disease (IBD).

EMAIL: jiaruilu@mail.med.upenn.edu

Longitudinal Data Analysis for Vaginal Microbiome Data

Eugenie M. Jackson*, University of Wyoming
Snehalata Huzurbazar, University of Wyoming

Human microbiome data is characterized by a high degree of sparseness and number of observations much smaller than the number of taxa, often with a small set of taxa dominating the microbial communities. Data are usually collected with the goal of relating microbial composition to covariates, especially, disease states. Increasingly, such data are collected longitudinally so that the need for analyzing these data with appropriate statistical techniques has become important. We present an overview of longitudinal techniques currently used in human, especially vaginal, microbiome studies. After discussing their strengths and drawbacks, we give an overview of our current work.

EMAIL: ejacks20@uwyo.edu

Integrative Analysis for Incorporating Human Microbiome to Improve Precision Treatment

Kyle M. Carter*, University of Arizona
Lingling An, University of Arizona

In recent years, the human microbiome (e.g. in gut), along with the host genome, has been discovered to play a critical role in disease progression and treatment effect. However, recent studies have also revealed that host genetic expression has a marked effect on the composition of species and associated functions in human microbiota. An integrative –omics study can be performed through mediation analysis, investigating how the human microbiome mediates the host gene expression on disease state, progression, treatment response, and more. Current high dimensional mediation models fail to incorporate the correlation between mediators (microbial species) or exposures (host genes) without data transformation which loses biological interpretation. In this research we propose a novel feature selection approach using a model-based entropy criterion in a multivariate zero-inflated negative binomial model to allow for correlated microbial mediators for small sample studies. Mediator microbes genes are selected in an iterative optimization algorithm. Through a series of comprehensive simulation studies, the proposed method shows superior performance to current methods.

EMAIL: kmcarter91@email.arizona.edu

Mediation Analysis of High-Dimensional Microbiome and Host Genome Data

Meng Lu*, University of Arizona
Lingling An, University of Arizona

Microbiome can have a major impact on the phenotype of their host, e.g., human health. Recent researches have been focused on association between human microbiota and the related diseases. In this research we are interested in understanding how the microbiome and host genome jointly impact human health and disease by integrating multiple –omics technologies through mediation analysis. Human genome is treated as mediator and microbiome as exposure. Currently there is lack of analytic approaches to dealing with high-dimensional correlated mediators (e.g., some genes are correlated due to a biological pathway) without any data transformation. We propose a regularized multivariate regression approach to identify the association between microbiome, human genome, and human disease status under a high dimensional setting. Joint significance test is used to evaluate the significance of the multiple correlated mediators while controlling the family-wise error rate.

EMAIL: menglu@email.arizona.edu

A Hierarchical Bayesian Mixed-Effects Model for High-Dimensional Microbiome Data Analysis

Neal S. Grantham*#, North Carolina State University
Brian J. Reich, North Carolina State University
Kevin R. Gross, North Carolina State University

Analysis of high-dimensional microbiome data with environmental covariates remains an open problem in microbiome research. Standard ecological analyses quantify differences between microbiome compositions through a single metric, such as Bray-Curtis dissimilarity, whereas recent approaches model the full microbiome as arising from a Dirichlet-multinomial distribution. Neither approach is well-suited for analyzing microbiome data as a response variable in designed experiments, as the for-

mer makes too drastic a reduction in the data to be useful and the latter does not account for mixed effects. In this paper, we develop MIMIX (MIcrobiome MIXed effects), a hierarchical Bayesian mixed-effects model for capturing per-taxon heterogeneity in the microbiome due to fixed and random effects in the experimental design. MIMIX offers a global test of treatment effect and local estimation of these effects on individual taxa while accounting for alternate sources of microbiome variability. We demonstrate the model on a 2x2 factorial experiment of the effects of fertilizer and herbivore exclusion on the fungal microbiome of a host crop species as part of the global Nutrient Network.

EMAIL: ngranth@ncsu.edu

22. Survival Analysis and Semi-Parametric and Non-Parametric Models

› Adjustment of Multi-Sample U-Statistics to Right Censored Data with Confounding Covariates

Yichen Chen*, University of Florida
Somnath Datta, University of Florida

We adjusted multi-sample U-statistics for comparison group distributions with right censored and confounding covariates by inverse probability of censoring weighting and score adjustment. The group membership, outcomes and censoring may depend on the covariates. Simulation results used to illustrate the performances of our U-statistics. In addition, Jackknife variance estimations were obtained and compared to the empirical values. A real data set on Bone Marrow transplant for Leukemia was applied to exemplify its usefulness.

EMAIL: yichen2014@ufl.edu

› Landmark Varying Coefficient Models for Predicting Long-Term Survival using Longitudinal Predictors in the Presence of Early Drop-Out

Jing Liu*, University of Florida
Xiaomin Lu, University of Florida

Most phase II oncology studies are conducted with an early endpoint observed within a short time period. Most such endpoints are based on underlying continuous covariates measured repeatedly over time. The endpoint itself is however categorized (of another continuous covariate) such as at a predefined time point or the change of covariates between time points which only uses partial information. Another common issue in oncology trials relates to missing data arising from early drop-outs. If patients drop out before the planned evaluation time for the phase II endpoint, they are either excluded or simply considered as failures for such endpoint, thus leading to a biased efficacy evaluation. To address these concerns, we propose two landmark time-varying predictive models utilizing all available longitudinal covariates including data from early-drop-out patients. In this work, we relax the independent censoring assumption to censoring at random. The predictive accuracy of the proposed models with longitudinal predictors are assessed and compared with traditional survival models with the conventional phase II outcome as predictors through data-based simulation studies.

EMAIL: liux1510@phhp.ufl.edu

› Induced Smoothing for Rank-Based Regression with Recurrent Gap Time Data

Tianmeng Lyu*, University of Minnesota
Xianghua Luo, University of Minnesota
Gongjun Xu, University of Minnesota
Chiung-Yu Huang, Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins University

A broad class of semiparametric regression models has recently been proposed for the analysis of gap times between consecutive recurrent events. Among them, the semiparametric accelerated failure time (AFT) model is especially appealing due to its direct interpretation. However, estimating regression coefficients for the AFT model could be challenging since its rank-based estimating function is a step function. Moreover, the popular resampling-based variance estimation for the AFT model requires solving rank-based estimating equations repeatedly and hence can be computationally inefficient. In this paper, we extend the induced smoothing method for the univariate AFT model to the case of recurrent gap time data. Our proposed smooth estimating function permits the application of standard numerical methods for the estimation of regression coefficients. Large sample properties and an asymptotic variance estimator are provided for the proposed method.

Simulation studies show that the proposed smooth method performs as well as the non-smooth method in terms of the regression coefficient estimates but with more efficient and computational reliable variance estimates.

EMAIL: lyuxx025@umn.edu

› Semiparametric Regression Analysis of Interval-Censored Data with Informative Dropout

Fei Gao*#, University of North Carolina, Chapel Hill
Donglin Zeng, University of North Carolina, Chapel Hill
Danyu Lin, University of North Carolina, Chapel Hill

Interval-censored data arise when the event of interest can only be ascertained through periodic examinations. In medical studies, subjects may not complete the examination schedule for reasons related to the event of interest. We develop a semiparametric approach to adjust for such informative dropout in the regression analysis of interval-censored data. Specifically, we propose a broad class of joint models, under which the event of interest and the dropout time follow transformation models with a shared random effect. We consider nonparametric maximum likelihood estimation and develop an EM algorithm with simple and stable calculations. We establish that the resulting estimators of the regression parameters are consistent and asymptotically efficient with a covariance matrix that can be consistently estimated by profile likelihood. We also show how to consistently estimate the survival function when dropout represents voluntary withdrawal versus an unavoidable terminal event. We assess the performance of the proposed methods through extensive simulation studies and provide an application to data on the diabetes incidence derived from a major epidemiological study.

EMAIL: fgao@live.unc.edu

› Transformation Model Estimation of Survival under Dependent Truncation and Independent Censoring

Sy Han Chiou*, Harvard School of Public Health
Matthew Austin, Harvard School of Public Health
Jing Qian, University of Massachusetts, Amherst
Rebecca Betensky, Harvard School of Public Health

Truncation is a mechanism that permits observation of selected subjects from a source population; subjects are included only if their event times are contained within subject-specific intervals. Standard survival analysis methods for estimation of the distribution of the event time require quasi-independence (Tsai, 1990) of failure and truncation. In the presence of quasi-dependence, alternative estimation procedures are required. We first consider two natural extensions of the transformation approach of Efron and Petrosian (1994) that handle right censoring and explain that these two methods generally do not produce a consistent estimator for the failure time distribution. To address this, we present a third extension that is consistent when the transformation model is approximately correct. A graphical diagnostic for assessment of the appropriateness of our proposed model is also proposed. We evaluate the proposed transformation model in simulations and apply it to the National Alzheimer Coordinating Centers autopsy cohort study, the Channing House data set and an AIDS incubation study.

EMAIL: schiou@hsph.harvard.edu

23. Clinical Trials and Adaptive Design/ Adaptive Randomization

› A Predictive Probability Interim Design for Phase II Clinical Trials with Continuous Endpoints

Meng Liu*, University of Kentucky
Emily Van Meter Dressler, University of Kentucky

Recent researches on targeted therapies have shifted drug development paradigm into establishing biological activity and target modulation in early phase trials. These trials need to address simultaneous evaluation of safety, proof-of-concept biomarker activity or changes in continuous tumor size instead of binary response rate. There is a lack of interim strategies developed to monitor futility or efficacy for these continuous outcomes, especially in single-arm phase II trials. We extend the design based on predictive probability proposed by Lee and Liu (2008) into a two-stage setting for continuous endpoints, assuming normal distribution with known variance. Simulation results and case study demon-

strated the proposed design can incorporate an interim stop for futility while maintaining desirable design properties for both optimal and minimax design, with reduced sample size. A limited exploration of priors was performed and shown to be robust. As research rapidly moves to incorporate more targeted therapies, it will accommodate new types of outcomes while allowing for flexible stopping rules to continue optimizing trial resources with compelling early phase data.

EMAIL: meng.liu@uky.edu

› A Bayesian Adaptive Design to Identify MTD Accounting for the Schedule Effects Using Pharmacodynamic Pharmacokinetic Modeling

Xiau Su*, University of Texas School of Public Health

In this study, we proposed a novel Bayesian adaptive dose-finding design to identify maximum tolerated dose for single schedule or multiple schedules accounting for the schedule effects. The basic idea is extending the traditional dose-response relationship into dose/schedule-concentration-response process using pharmacodynamic-pharmacokinetic. Therefore, the schedule information as well as the drug concentration data can be utilized. The information can be borrowed across different schedules for multiple-schedule problems. The simulation studies show that the proposed design outperforms CRM, BACRM, mTPI and nonparametric optimal design in most scenarios.

EMAIL: xsu2@mdanderson.org

› An Adaptive Multi-Stage Phase I Dose-Finding Design Incorporating Continuous Efficacy and Toxicity Data from Multiple Treatment Cycles

Yu Du*, Johns Hopkins University
Jun Yin, Mayo Clinic
Daniel Sargent, Mayo Clinic
Sumithra Mandrekar, Mayo Clinic

Phase I designs traditionally use the dose-limiting toxicity (DLT), a binary endpoint based on the first treatment cycle, to identify the maximum-tolerated dose (MTD) assuming a monotonic relationship between dose and efficacy. In this article, we propose a multi-stage adaptive Bayesian approach by jointly modeling an efficacy outcome and toxicity endpoints from multiple treatment cycles. We replace DLT with the normalized total toxicity profile (nTTP), a quasi-continuous endpoint, to take into account clinical multidimensionality (multiple grades and types) of toxicities. In addition, our design accommodates longitudinal toxicity data in order to capture the adverse events from multiple cycles, and non-monotone dose-efficacy relationships. Simulations show that the design has a high probability of making the correct dose selection and has good overdose control across various scenarios. Our design terminates early when all doses are toxic. To our best knowledge, the proposed design is the first to incorporate multiple cycles of toxicities and a continuous efficacy outcome.

EMAIL: duyu8411@gmail.com

› Hypothesis Testing of Adaptive Randomization to Balance Continuous Covariates

Xiaoming Li*, Johnson and Johnson
Jianhui Zhou, University of Virginia
Feifang Hu, The George Washington University

Covariate-adaptive designs are widely used to balance covariates and maintain randomization in clinical trials. However, the understanding of adaptive designs with continuous covariates lacks a theoretical foundation. Herein, we establish a theoretical framework for hypothesis testing on adaptive designs with continuous covariates based on linear models. We test for treatment effects and significance of covariates under null and alternative hypotheses. To verify our theoretical framework, numerical simulations are conducted under a class of covariate-adaptive designs. Key findings using independent covariates based on linear models include: (1) hypothesis testing that compares treatment effects is conservative in terms of smaller type I error, (2) hypothesis testing to compare treatment effects using adaptive designs outperforms complete randomization method in terms of power, and (3) testing for significance of covariates is still valid.

EMAIL: xl3xa@virginia.edu

› Adaptive Dose Modification for Phase I Clinical Trials

Yiyi Chu*, University of Texas School of Public Health
Haitao Pan, University of Texas MD Anderson Cancer Center
Ying Yuan, University of Texas MD Anderson Cancer Center

Most phase I dose-finding methods in oncology aim to find the maximum-tolerated dose (MTD) from a set of prespecified doses. However, in practice, due to a lack of understanding of the true dose-toxicity relationship, it is likely that none of these prespecified doses is equal or reasonably close to the true MTD. To handle this issue, we propose an adaptive dose modification (ADM) method that can be coupled with any existing dose-finding method to adaptively add a dose, when it is needed, during the course of dose finding. To reflect clinical practice, we divide the toxicity probability into three regions: underdosing, acceptable and overdosing regions. We adaptively add a dose whenever the observed data suggest that none of the investigational doses is likely to be located in the acceptable region. The added dose is estimated based on local polynomial regression. The simulation study shows that ADM outperforms the similar existing method, with more precise added doses. We applied ADM to a phase I cancer trial.

EMAIL: yiyi.chu@uth.tmc.edu

› MIDAS: A Practical Bayesian Design for Platform Trials with Molecularly Targeted Agents

Ying Yuan*, University of Texas MD Anderson Cancer Center

Recent success of immunotherapy and other targeted therapies in cancer treatment has led to an unprecedented surge in the number of novel therapeutic agents that need to be evaluated in clinical trials. Traditional phase II clinical trial designs were developed for evaluating one candidate treatment at a time, and thus not efficient for this task. We propose a Bayesian phase II platform design, the Multi-candidate Iterative Design with Adaptive Selection (MIDAS), which allows investigators to continuously screen a large number of candidate agents in an efficient and seamless fashion. MIDAS consists of one control arm, which contains a standard therapy as the control, and several experimental arms, which contain the experimental agents. During the trial, we adaptively drop inefficacious or overly toxic agents and graduate the promising agents from the trial to the next stage of development. Whenever an experimental agent graduates or is dropped, the corresponding arm opens immediately for testing the next available new agent. Simulation studies show that MIDAS substantially outperforms the conventional approach.

EMAIL: yyuan@mdanderson.org

24. Medical Device Applications

› Passing-Bablok Regression Analysis with Rank-Transferred Data in Evaluating Hemostasis State of a Blood Sample

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

Hemostasis evaluations are commonly used to assess clinical conditions in trauma, cardiovascular surgery and cardiology procedures to assess hemorrhage or thrombosis conditions before, during and following the procedure. Clotting time, clot stiffness/firmness, and platelet function are parameters to be evaluated when the new device is seeking a substantial equivalence to an already marketed device. Although these parameters are semi-quantitative, if the units are different between the investigational and the predicate devices due to using different techniques to measure these parameters, regression analyses are not an appropriate method to use. In this talk, whether using Passing-Bablok regression analysis with rank-based transferred data may be the possible analysis choice will be reviewed with a simulated data set. Some statistical challenges and issues in study design, and sample size calculation will be discussed as well.

EMAIL: kyungsook.kim@fda.hhs.gov

› Overall Unscaled Indices for Quantifying Agreement Among Multiple Raters

Jeong Hoon Jang*, Emory University
Amita Manatunga, Emory University
Qi Long, Emory University

The need to analyze agreement exists in various study fields where quantifying inter-rater reliability is of great importance. Several unscaled agreement indices, such as total devia-

tion index (TDI) (Lin 2000) and coverage probability (CP) (Lin et al. 2002) are widely recognized for two reasons: (1) they are intuitive in a sense that interpretations are tied to the original measurement unit; (2) practitioners can readily determine whether the agreement is satisfactory by directly comparing the value of the index to a pre-specified tolerable coverage probability or difference. However, these indices were only defined in the context of comparing two raters. In this presentation, we introduce a series of extended unscaled indices that can be used to evaluate agreement among multiple raters. We derive the definitions of extended indices and propose inference procedures in which bootstrap methods are used for the estimation of standard errors. We assess the performance of the proposed approaches by simulation studies. Finally, we demonstrate the application of our methods via application to a renal study.

EMAIL: jjang54@emory.edu

› Reference Database for In Vivo Diagnostic Devices

Bipasa Biswas*, U.S. Food and Drug Administration

Clinical laboratory tests often require a reference interval for quantitative tests and the construction of such intervals are common for laboratory tests. In vivo diagnostic devices in ophthalmology or neurology utilize reference database (also commonly known as normative database) to compare an individual patient's medical device output/result against the reference database for clinical management of the individual patient. This talk will discuss issues related to subject selection, important co-variate, presentation and use of results from a reference database.

EMAIL: bipasa.biswas@fda.hhs.gov

› A Two-stage Model for Wearable Device Data

Jiawei Bai*#, Johns Hopkins University
Yifei Sun, Johns Hopkins University
Jennifer A. Schrack, Johns Hopkins University
Ciprian M. Crainiceanu, Johns Hopkins University
Mei-Cheng Wang, Johns Hopkins University

Recent advances of wearable computing technology have allowed continuous health monitoring in large observational studies and clinical trials. Examples of data collected by wearable devices include minute-by-minute physical activity proxies measured by accelerometers or heart rate. The analysis of data generated by wearable devices has so far been quite limited to crude summaries, for example, the mean activity count over the day. To better account for the temporal and population variability, we introduce a two-stage regression model for the minute-by-minute physical activity proxy data. Two approaches were proposed to estimate both time-varying parameters and structural parameters. The model is designed to capture both the transition dynamics between active/inactive periods (Stage 1) and activity intensity dynamics during active periods (Stage 2). The approach extends methods developed for zero-inflated Poisson data to account for the high-dimensionality and time-dependence of the high density data generated by wearable devices. Methods are motivated by and applied to the Baltimore Longitudinal Study of Aging.

EMAIL: javybai@gmail.com

› Generalized Linear Mixed Models for Analysis of Cross-Correlated Binary Response in Multireader Studies of Diagnostic Accuracy

WITHDRAWN

Yuvika Paliwal*, University of Pittsburgh
Andriy I. Bandos, University of Pittsburgh

Generalized Linear Mixed Models (GLMMs) provide a powerful built-in tool for analyzing cross-correlated data, especially in presence of covariates. However, conventional GLMMs for binary data could lead to biased results. Yet, the quality of statistical inferences are generally unknown for data typically encountered in multireader studies of diagnostic accuracy (where each of several readers evaluates the same sample of subjects for presence/absence of a specific condition). Furthermore, the structure of these studies favors consideration of much less known half-marginal models that may have better properties due to substantially reducing the number of random effects by marginalizing over subjects (e.g., for estimating sensitivity). We investigated GLMMs for cross-correlated binary data with and without covariates. Using simulations we evaluated estimation bias as well as the coverage of the confidence intervals for the fixed effects. We demonstrated that in common multireader scenarios more standard subject-specific models often result in bias and degraded

CI coverage, whereas the better-interpretable half-marginal models showed low bias and adequate coverage of the CIs.

EMAIL: yup3@pitt.edu

› Assessing Non-Inferiority Based on Risk Difference in Non-Randomized, One-to-Many Matched Studies

Jeremiah Perez*, Boston University
Joseph Massaro, Boston University

Non-inferiority (NI) tests are well developed for randomized parallel group trials where the control and experimental groups are independent. However, these tests may not be appropriate for assessing NI in matched study designs where many control subjects are matched to each experimental subject. These tests may require an adjustment to account for the correlation among matched subjects. We propose a new statistical test that extends Farrington-Manning (FM) test to the case of correlated one-to-many matched data. We conducted a simulation study to compare the size and power of the proposed test statistic with tests developed for clustered matched pair data. In the presence of intra-class correlation, the sizes of tests developed for clustered matched pair data are inflated when applied to one-to-many matched data. The size of the proposed method, on the other hand, is close to the nominal level for a variety of correlation patterns. As the correlation between the experimental and control groups increases, the proposed test becomes more powerful than FM test. The proposed method is a good alternative to FM test for assessing NI in one-to-many matched study designs.

EMAIL: jperez1@bu.edu

25. Methods for Single-Cell Analysis

› Accounting for Technical Noise in Cell Type Identification by Single-Cell RNA-Sequencing

Rong Ma*, University of Pennsylvania
Nancy Zhang, University of Pennsylvania
Mingyao Li, University of Pennsylvania

Recent development of single-cell RNA-seq (scRNA-seq) has led to enormous biological discoveries yet also introduced statistical challenges. An important step in scRNA-seq analysis is to identify cells belonging to the same cell types based on their gene expression. However, technical noise introduced during scRNA-seq experiments can obscure cell type identification if not appropriately controlled. To tackle this problem, we propose a three-step procedure IceT (Identification of cell Types using Recovered gene expression). In step 1, IceT "recovers" the distribution of true expression of a gene among cells using ERCC spike-ins. In step 2, IceT employs nonparametric density estimation to select genes informative for cell type identification. In step 3, the selected genes are used to identify cell types with the aid of tSNE. Through simulations, we show that the denoised gene expression obtained from the recovery procedure enable reliable selection of informative genes, which allows us to better classify cell types when compared to methods relying on observed gene expression. We further apply IceT to a mouse cortex data and show that it is able to refine cell types.

EMAIL: rongm@mail.med.upenn.edu

› Statistical Analysis of Single-Cell Chromatin Accessibility Data (scATAC-seq)

Peter F. Hickey*, Johns Hopkins University
Kasper D. Hansen, Johns Hopkins University

New single-cell genomic technologies are generating data from individual cells on a diverse set of genomic measurements. These assays offer enormous promise to researchers seeking to understand cellular heterogeneity, to identify and characterise novel cell types, and to study rare subpopulations of cells, with application to cancer research, developmental biology, and much more. The single-cell assay for transposase-accessible chromatin using sequencing (scATAC-seq) can be used to measure the chromatin accessibility from individual cells. The data from each cell are very sparse, have a limited dynamic range (typically, only 0, 1, or 2 reads may be observed at any position in the genome), and are zero-inflated. Additionally, the number of cells profiled is small (in the tens or hundreds), and like all single-cell genomics assays, scATAC-seq data display a large amount of cell-to-cell heterogeneity and technical artefacts that may confound biological signals. We will present

our work addressing some of these challenges in the statistical analysis of scATAC-seq data.

EMAIL: peter.hickey@gmail.com

› Accounting for Technical Noise in Single-Cell RNA Sequencing Analysis

Cheng Jia*, University of Pennsylvania
Derek Kelly, University of Pennsylvania
Junhyong Kim, University of Pennsylvania
Mingyao Li, University of Pennsylvania
Nancy R. Zhang, University of Pennsylvania

Recent technological breakthroughs have made it possible to measure RNA expression at the single-cell level, paving the way for exploring expression heterogeneity among individual cells. Current single-cell RNA sequencing (scRNA-seq) protocols introduce technical biases that vary across cells, and can bias downstream analysis. To adjust for cell-to-cell technical differences, we propose a statistical framework, TASC (Toolkit for Analysis of Single Cell RNA-seq), an empirical Bayes approach to model the cell-specific dropout rates and amplification efficiency by use of external RNA spike-ins. TASC incorporates the estimated technical parameters, which reflect cell-to-cell batch effects, into a hierarchical mixture model to estimate the biological variance of a gene, detect differentially expressed (DE) genes, and test for transcriptional bursting. TASC is also able to adjust for additional covariates to further eliminate potential confounding. In our simulations based on real scRNA-seq data, TASC displays superior sensitivity and specificity in detection of DE genes. We believe that TASC will provide a robust platform for researchers to leverage the power of scRNA-seq.

EMAIL: jiacheng@mail.med.upenn.edu

› A Novel Method for Analyzing Single Cell Transcriptomic Data

Zhe Sun*, University of Pittsburgh
Ming Hu, Cleveland Clinic
Wei Chen, Children's Hospital of Pittsburgh of UPMC and University of Pittsburgh

Single cell transcriptome sequencing technology can uncover cell heterogeneity in gene expression and has become a revolutionary tool to study molecular processes within complex cell population. We have been using the recently released droplet-based Chromium system from 10X Genomics to study cell heterogeneity of tissues collected from human and mouse samples. Despite technology advance, statistical methods and computational tools are still lacking for clustering droplet-based single cell transcriptomic data. Most of existing methods, such as K-means and hierarchical clustering, are deterministic, thus cannot provide quantification of statistical uncertainty. More importantly, no method explicitly characterizes multiple layers of uncertainties from multiple sources due to intrinsic data characteristics. We developed a mixture model to characterize cell types and understand cell heterogeneity based on single cell transcriptomic data. In both simulation and real data studies, our proposed method outperformed existing methods and provided extra information for downstream analysis. Our method will facilitate ongoing single cell studies and accelerate novel biological discovery.

EMAIL: zhs31@pitt.edu

› Detecting Cellular Heterogeneity Based on Single Cell Sequencing

WITHDRAWN

Kyounga Song*, University of Arizona
Lingling An, University of Arizona

Single cell sequencing analysis allows us to obtain the heterogeneous information of individual cells, which is assumed homogeneous in bulk cell analysis. Cellular heterogeneity could play a key role in answering the question in cancer research, stem cell biology and immunology. Although technical variation can be adjusted by normalization, there still exist two types of biological variations, biological heterogeneity of interest and unwanted biological noises due to cell sizes and cell cycles. We propose an innovative statistical approach to capture wanted biological heterogeneity by gene-group structure that cells contain, while unwanted biological noises are greatly reduced. Gene-group structure is determined by a Dirichlet model, a nonparametric Bayesian method, on expression of all genes across all samples/cells under all involved treatments/conditions. Using simulated

data we demonstrate that the proposed method is able to capture the heterogeneity of interest in gene expression, with greater power in detection than existing methods in single cell sequencing analysis.

EMAIL: ksong@email.arizona.edu

Modeling Allele-Specific Gene Expression by Single-Cell RNA Sequencing

Yuchao Jiang*, University of Pennsylvania
Nancy R. Zhang, University of Pennsylvania
Mingyao Li, University of Pennsylvania

Allele-specific expression is traditionally studied by bulk RNA sequencing, which measures average expression across cells. Single-cell RNA sequencing (scRNA-seq) allows the comparison of expression distribution between the two alleles of a diploid organism, and characterization of allele-specific bursting. We propose SCALE to model genome-wide allele-specific bursting at the allelic level, while accounting for technical bias and other complicating factors such as cell size. SCALE detects genes with significantly different bursting kinetics between the two alleles, as well as genes where the two alleles exhibit dependence in their bursting processes. We illustrate SCALE on a mouse dataset and find that, globally, cis control in gene expression acts through modulation of burst frequency, and that a significant number of genes exhibit coordinated bursting between alleles. Through simulations, the effects of technical noise, cell size, and the true values of the bursting kinetic parameters on estimation accuracy are explored.

EMAIL: yuchaoj@wharton.upenn.edu

26. Epidemiologic Methods and Study Design

Calibration and Seasonal Adjustment for Matched Case-Control Studies of Vitamin D and Cancer

Mitchell H. Gail*, National Cancer Institute, National Institutes of Health
Jincao Wu, U.S. Food and Drug Administration
Molin Wang, Harvard School of Public Health
Shiaw-Shyuan Yaun, Harvard School of Public Health
Marjorie L. McCullough, American Cancer Society
Kai Yu, National Cancer Institute, National Institutes of Health
Anne Zeleniuch-Jacquotte, New York University School of Medicine
Stephanie A. Smith-Warner, Harvard School of Public Health
Regina G. Ziegler, National Cancer Institute, National Institutes of Health
Raymond J. Carroll, Texas A&M University and University of Technology, Sydney

Vitamin D measurements are influenced by seasonal variation and specific assay used. Motivated by multicenter studies of vitamin D with cancer, we give an analytic framework for matched case-control data that accounts for seasonal variation and calibrates to a reference assay. Key findings include: (1) Failure to adjust for season and calibration increased variance, bias and mean square error. (2) Analysis of continuous vitamin D requires a variance adjustment for variation in the calibration estimate, but log odds estimates do not depend on a reference date for seasonal adjustment. (3) For categorical vitamin D risk models, procedures based on categorizing the seasonally adjusted and calibrated vitamin D levels have near nominal operating characteristics. However, estimates of category-specific log odds ratios depend on the reference date for seasonal adjustment. Thus public health recommendations based on categories of vitamin D should also define the time of year to which they refer. This work is informing the analyses of the multicenter Vitamin D Pooling Project for Breast and Colorectal Cancer.

EMAIL: gailm@mail.nih.gov

The Impact of Informative Wear Time on Modeling Physical Activity Data Measured Using Accelerometers

Jaejoon Song*, University of Texas School of Public Health
Michael D. Swartz, University of Texas School of Public Health
Kelley Pettee Gabriel, University of Texas School of Public Health
Karen Basen-Engquist, University of Texas MD Anderson Cancer Center

Wearable sensors provide an exceptional opportunity in collecting real-time behavioral data in free living conditions. However, wearable sensor data from observational studies often suffer from information bias, since participants' willingness to wear the monitoring devices may be associated with the underlying behavior of interest. The aim of this study was to quantify bias in estimating associations between a fixed predictor and levels of physical activity, measured using wearable sensors. Our simulation study indicated that estimates from the conventional methods for longitudinal data analysis showed small or no bias when device wear patterns were independent of the participants' physical activity process, but biased against the null when the patterns of device non-wear times were associated with the physical activity process. We introduce a semiparametric statistical approach in modeling physical activity data from wearable sensors. The estimates from the semiparametric modeling approach were unbiased both when the device wear patterns were i) independent or ii) dependent to the underlying physical activity process.

EMAIL: jaejoonsong@gmail.com

Scale Weighted Zero-Inflated Negative Binomial Model for Count Data from a Complex Multilevel Survey

Lin Dai*, Medical University of South Carolina
Mulugeta Gebregziabher, Medical University of South Carolina

We demonstrate a novel application of a weighted zero-inflated negative binomial model to quantify regional variation in HIV-AIDS prevalence in sub-Saharan African countries. We use data from latest round of the Demographic and Health survey (DHS) conducted in three countries (Ethiopia-2011, Kenya-2009 and Rwanda-2010). The outcome is an aggregate count of HIV cases in each census enumeration area (CEA) from the DHS of the three sub-Saharan African countries. Data are characterized by excess zeros and heterogeneity due to clustering. We compare several scale-weighting approaches to account for the complex survey design and clustering in a zero inflated negative binomial (ZINB) model. Finally, we provide marginalized rate ratio (RR) estimates from the best ZINB model.

EMAIL: daili@musc.edu

Investigating the Assumptions of the Self-Controlled Case Series Method

Heather Whitaker, The Open University, UK
Yonas Ghebremicahel-Weldeselassie*, The Open University, UK
Ian Douglas, London School of Hygiene and Tropical Medicine
Liam Smeeth, London School of Hygiene and Tropical Medicine
Paddy Farrington, The Open University, UK

The self-controlled case series (SCCS) method is an epidemiological study design for which individuals act as their own control i.e. comparisons are made within-individuals. Hence, only individuals who have experienced an event are included and all time invariant confounding is eliminated. We describe some simple techniques for investigating two key assumptions of the self-controlled case series method, namely that events do not influence subsequent exposures, and that events do not influence observation periods. For each assumption we propose some simple tests based on the standard SCCS model, along with associated graphical displays. The methods also enable the user to investigate the robustness of the results obtained using the standard SCCS model to failure of assumptions. The proposed methods are investigated by simulations, and applied to data on measles, mumps and rubella vaccine, and antipsychotics.

EMAIL: yonas.weldeselassie@open.ac.uk

› Novel Statistical Framework to Study Fragmentation of Daily Physical Activity

Junrui Di*, Johns Hopkins Bloomberg School of Public Health
Andrew Leroux, Johns Hopkins Bloomberg School of Public Health
Jacek Urbanek, Johns Hopkins Bloomberg School of Public Health
Jennifer Schrack, Johns Hopkins Bloomberg School of Public Health
Adam Spira, Johns Hopkins Bloomberg School of Public Health
Vadim Zipunnikov, Johns Hopkins Bloomberg School of Public Health

Sedentary behavior is a significant risk factor for a wide range of chronic diseases, comorbidities, and mortality. As people age, daily activities become less fragmented and sedentary behaviors increase. Existing methods of studying sedentary behaviors often rely on quantifying total sedentary volume and ignore the accumulation and temporal distribution of sedentary time. To deal with this inadequacy, we developed a unifying statistical framework to study the fragmentation of physical activity measured with accelerometry by analyzing the distribution functions of sedentary and active bout durations using both parametric and nonparametric approaches. We illustrate the approach by exploring the association of fragmentation of physical activity and mortality in National Health and Nutrition Examination Survey (NHANES) 2003-2006.

EMAIL: jdi2@jhu.edu

› Enhancing Power of Case-Control Studies by Using Prevalent Cases

Marlena Maziarz*, National Cancer Institute, National Institutes of Health
Jing Qin, National Institute of Allergy and Infectious Diseases, National Institutes of Health
Ruth Pfeiffer, National Cancer Institute, National Institutes of Health

When assessing associations of exposures with rare diseases based on case control studies designed within well-defined cohorts, individuals diagnosed prior to cohort entry are typically excluded to avoid the potential impact of survival bias on study findings. We developed methods that in addition to data on controls and incident cases allow one to include information from prevalent cases to improve efficiency of case-control studies. We construct a constrained empirical likelihood assuming an exponential tilting model that leads to logistic regression and obtain efficient estimates of association parameters. We adjust for survival bias by modeling the backward time for prevalent cases using a parametric survival distribution. We develop an empirical likelihood ratio test for the association parameters in the logistic or survival model. We quantify the efficiency gain when incident cases are supplemented with prevalent cases in simulations, and illustrate our methods by estimating the association of single nucleotide polymorphisms (SNPs) with breast cancer risk based on data from the U.S. Radiologic Technologists Health Study, a prospective cohort of radiologic technologists.

EMAIL: marlena.maziarz@nih.gov

27. New Advances in Biomarker Evaluation and Predictive Modeling

› Preconditioning Method for Development of Risk Prediction Tools

Dandan Liu*, Vanderbilt University
Cathy A. Jenkins, Vanderbilt University
Frank E. Harrel, Vanderbilt University

With emerging trend of large-scale data resources, novel methods of risk prediction modeling is advocated, yet challenging. Although modern feature selection methods for high-dimensional data have been developed for decade, they have not been sufficiently translated into development of risk prediction tools which additionally involves prediction on top of variable selection. In this talk, a preconditioning method for risk prediction modeling will be discussed. This two-step method separates risk prediction from model selection and thus is both flexible in terms of the choice of feature selection methods and stable in terms of the final selected model. In the first step, a preconditioned outcome is obtained utilizing the correlation

structure of the candidate risk factors. In the second step, model selection is performed against the preconditioned outcome yielding the final risk prediction model. A risk prediction tool for 30-day adverse event of AHF patients admitted to ED was developed using the proposed method.

EMAIL: Dandan.Liu@Vanderbilt.edu

› Methods for Incorporating Large-Scale Sparsely-Measured Longitudinal Biomarkers in Dynamic Hazards Models

Yuanjia Wang*, Columbia University
Xiang Li, Columbia University
Qufeng Li, University of North Carolina, Chapel Hill
Donglin Zeng, University of North Carolina, Chapel Hill
Karen Marder, Columbia University

Clinical studies of time-to-event outcome often collect a large number of dynamic covariates over time to build time-sensitive prognostic model and inform individualized treatment. Due to the resource-intensive and complex data collection process, large-scale biomarkers may be collected infrequently and thus not available at every event time point. Estimating time-varying effects of high-dimensional dynamic biomarkers infrequently measured is a challenging problem. We propose a time-varying coefficient hazards model, where we apply kernel smoothing to borrow information across subjects and among biomarkers to remedy sparse and unbalanced assessment time for each individual and propose a penalized pseudo-likelihood function for estimation. An efficient algorithm inspired by the alternating direction method of multipliers is adopted for computation. Under some regularity conditions, we show that with ultra-high dimensional covariates, the proposed method enjoys oracle property. Through extensive simulation studies and real data applications, we demonstrate superior performance in terms of estimation and selection.

EMAIL: yw2016@cumc.columbia.edu

› Dynamic Predictions with Two-Phase Study Designs

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

Little attention has been given to the design of rigorous and efficient studies to evaluate longitudinal biomarkers. Measuring longitudinal markers on an entire cohort is cost prohibitive and, especially for rare outcomes such as cancer, may be infeasible. Thus, methods for evaluation of longitudinal biomarkers using efficient and cost-effective study designs are needed. Case-cohort (CCH) and nested case-control (NCC) studies allow investigators to evaluate biomarkers rigorously and at reduced cost, with only a small loss in precision. We develop estimators of several measures to evaluate the accuracy and discrimination of predicted risk under CCH and NCC study designs. To facilitate inference with under two-phase longitudinal data, we develop valid resampling-based variance estimation procedures under CCH and NCC. These methods are widely applicable, efficient and cost-effective, and can be easily adapted to other study designs used to evaluate prediction rules in a longitudinal setting.

EMAIL: yzheng@fhcrc.org

› Using Longitudinal Biomarker Data to Dynamically Predict Time to Disease Progression

Xuelin Huang*, University of Texas MD Anderson Cancer Center
Fangrong Yan, China Pharmaceutical University
Ruosha Li, University of Texas Health Science Center at Houston
Jing Ning, University of Texas MD Anderson Cancer Center
Ziding Feng, University of Texas MD Anderson Cancer Center

New statistical methods are proposed for conducting predictions on a real-time basis so that at any time during follow-up, as soon as a new biomarker value is obtained, the predictions can be updated immediately to reflect the latest prognosis. These methods include quantile regression on residual survival time, functional principal component analysis for summarizing the changing patterns of patients' longitudinal biomarker trajectories, and a supermodel to smoothly extend landmark analyses on discrete time points to the whole follow-up time interval. Simulation studies show that the proposed approaches achieve stable estimation of biomarker effects over time, and are robust against model misspecification. Moreover, they have better predictive performance than current methods, as evaluated by the root

mean square error and area under the curve of receiver's operating characteristics.

EMAIL: xlhuang@mdanderson.org

28. Statistical Methods for Human Microbiome Data Analysis

› A General Framework for Association Analysis of Microbial Communities on a Taxonomic Tree

Zheng-Zheng Tang*, Vanderbilt University School of Medicine
Guanhua Chen, Vanderbilt University School of Medicine
Alexander V. Alekseyenko, Medical University of South Carolina
Hongzhe Li, University of Pennsylvania

Recent advances in high-throughput sequencing technology have made it possible to obtain data on the composition of microbial communities and to study the effects of dysbiosis on the human host. We develop a general framework to: (a) perform robust association tests for the microbiome data that exhibit arbitrary inter-taxa dependencies; (b) localize lineages on the taxonomic tree that are associated with covariates (e.g. disease status); and (c) assess the overall association of the microbial community with the covariates of interest. Unlike existing methods for microbiome association analysis, our framework does not make any distributional assumptions on the microbiome data; it allows for the adjustment of confounding variables and accommodates excessive zero observations; and it incorporates taxonomic information. We performed extensive simulation studies under a wide-range of scenarios to evaluate the new methods and showed substantial power gain over existing methods. The advantages of the proposed framework are further demonstrated with real datasets from two microbiome studies.

EMAIL: z.tang@vanderbilt.edu

› Composition Estimation from Sparse Count Data via a Regularized Likelihood

Yuanpei Cao, University of Pennsylvania
Anru Zhang, University of Wisconsin, Madison
Hongzhe Li*, University of Pennsylvania

In microbiome studies, taxa composition is often estimated based on the sequencing read counts in order to account for the large variability in the total number of observed reads across different samples. Due to sequencing depth, some rare microbial taxa might not be captured in the metagenomic sequencing, which results in many zero read counts. Naive composition estimation using count normalization therefore lead many zero proportions, which underestimates the underlying compositions, especially for the rare taxa. In this paper, the observed counts are assumed to be sampled from a multinomial distribution, with the unknown composition being the probability parameter in a high dimensional positive simplex space. Under the assumption that the composition matrix is approximately low rank, a nuclear norm regularization-based likelihood estimation is developed to estimate the underlying compositions of the samples. Simulation studies demonstrate that the regularized maximum likelihood estimator outperforms the commonly used naive estimators. The methods are applied to an analysis of a human gut microbiome dataset.

EMAIL: hongzhe@upenn.edu

› Models for Microbial Community Population Structure

Boyu Ren, Harvard School of Public Health
Emma Schwager, Harvard School of Public Health
Jason Lloyd-Price, Harvard School of Public Health
Steffen Ventz, University of Rhode Island
Eric A. Franzosa, Harvard School of Public Health
Curtis Huttenhower*, Harvard School of Public Health

Culture-independent microbial community sequencing data present challenges during analysis due in part to their quantitative properties. They are typically noisy, sparse (zero-inflated), high-dimensional, and extremely non-normal, often arising in the form of either count or compositional measurements. I will discuss Bayesian models for two common types of microbiome data, taxonomic profiles (which indicate the abundances of organismal features) and ecological interactions (i.e. significant co-occurrences or co-variation). The first, SparseDOSSA, parameterizes typical microbial sequencing counts across taxa and samples and allows, in turn, realistic simulated data generation for methods development. The second, BAnOCC, infers covariance between unobserved basis features (i.e. absolute

microbial abundance measurements) given compositional data (which is generated by typical sequencing data). I will conclude with comments on other types of graphical models (e.g. Gaussian processes) for microbiome epidemiology studies.

EMAIL: chuttenh@hsph.harvard.edu

› Deflating Error and Increasing Power in Permutational Analysis of Variance on Distances with Heteroscedastic Microbiome Data

Alexander V. Alekseyenko*, Medical University of South Carolina

Mirobiome data comes as multivariate abundance profiles, which are analyzed by the microbial ecologists in terms of pairwise distances between observations, called beta diversity measures. Suppose observations are grouped into k discrete categories. Inference about association of the grouping with the microbiota can be obtained using non-Euclidean permutational multivariate analysis of variance (PERMANOVA). We have previously shown that in the two-sample case PERMANOVA exhibits loss of power or inflated type I error in presence of heteroscedasticity and unbalanced samples. We have derived multivariate Welch T-test on distances to overcome this issue. The resulting permutational test has the desired type I error and power properties. In this work, we extend the solution to arbitrary number of categories. We do so by considering heteroscedastic Welch ANOVA statistic W* and deriving distance-based calculation to extend it to the multivariate case. We show that our previous Welch T-test is a special case of this more general solution. We further empirically evaluate the performance of the distance-based multivariate W* permutation test, against several alternative strategies.

EMAIL: alekseye@musc.edu

29. Advances in Measurement Error Methods

› Generalized Method-of-Moments Estimation and Inference for the Validation of Multiple Imperfect Measures

Donna Spiegelman*, Harvard School of Public Health
Samuela Pollack, Harvard School of Public Health
Raymond J. Carroll, Texas A&M University
Xin Zhou, Harvard School of Public Health
Walter Willett, Harvard School of Public Health
Eric Rimm, Harvard School of Public Health

We develop semi-parametric generalized method of moments estimators for the regression calibration de-attenuation factor and other quantities, including the correlation of each surrogate measure with the unobserved truth and the intra-class correlation coefficients characterizing the random within-person variation around each measurement. The method makes assumptions only about the first two moments of the joint multivariate distribution of the data. The robust variance is derived to allow asymptotic inference. We consider partially and fully efficient iterative methods for this typically over-determined problem. Data from an unbiased gold standard and other objective and subjective measures are assumed available. Methods are applied to Harvards Womens Lifestyle Validation Study, which measured physical activity by doubly labeled water, accelerometer, pulse, questionnaire (PAQ), and ACT24, an on-line PA assessment tool. Using these 5 measures, the correlations of PAQ and ACT24 with truth were 0.42 (0.38, 0.46) and 0.27 (0.23, 0.31), and the correlations of the accelerometer and resting pulse with truth were 0.73 (0.72, 0.74) and -0.24 (-0.13, -0.35).

EMAIL: stdls@hsph.harvard.edu

› Proportional Hazards Model with Covariate Measurement Error and Instrumental Variables

Xiao Song*, University of Georgia
Ching-Yun Wang, Fred Hutchinson Cancer Research Center

In biomedical studies, covariates with measurement error may occur in survival data. Existing approaches mostly require cer-

tain replications on the error-contaminated covariates, which may not be available in the data. In this paper, we develop a simple nonparametric correction approach for estimation of the regression parameters in the proportional hazards model using a subset of the sample where instrumental variables are observed. The instrumental variables are related to the covariates through a general nonparametric model, and no distributional assumptions are placed on the error and the underlying true covariates. We further propose a novel generalized methods of moments nonparametric correction estimator to improve the efficiency over the simple correction approach. The efficiency gain can be substantial when the calibration subsample is small compared to the whole sample. The estimators are shown to be consistent and asymptotically normal. Performance of the estimators is evaluated via simulation studies and by an application to data from an HIV clinical trial.

EMAIL: xsong@uga.edu

› A Class of Semiparametric Tests of Treatment Effect Robust to Confounder Classical Measurement Error

Caleb H. Miles*, University of California, Berkeley
Joel Schwartz, Harvard School of Public Health
Eric J. Tchetgen Tchetgen, Harvard School of Public Health

When assessing the presence of a particular causal effect, it is well known that classical measurement error of the exposure can reduce the power of a test of the null hypothesis in question, although its type I error rate will generally remain at the nominal level. In contrast, classical measurement error of a confounder can inflate the type I error rate of a test of treatment effect. In this paper, we develop a large class of semiparametric test statistics of a causal effect, which are completely robust to classical measurement error of a subset of confounders. A unique and appealing feature of our proposed methods is that they require no external information such as validation data or replicates of error-prone confounders. We present a doubly-robust form of this test that requires only one of two models to be correctly specified for the resulting test statistic to have correct type I error rate. We demonstrate validity and power within our class of test statistics through simulation studies. We apply the methods to a multi-U.S.-city, time-series data set to test for an effect of temperature on mortality while adjusting for PM2.5, which is known to be measured with error.

EMAIL: calebhmiles@gmail.com

› Variable Selection and Inference Procedures for Marginal Analysis of Longitudinal Data with Missing Observations or Measurement Error

Grace Y. Yi*, University of Waterloo
Xianming Tan, University of North Carolina, Chapel Hill
Runze Li, The Pennsylvania State University

In contrast to extensive attention on model selection for univariate data, research on correlated data remains relatively limited. Furthermore, in the presence of missing data and/or measurement error, standard methods would typically break down. To address these issues, we propose marginal methods that simultaneously carry out model selection and estimation for longitudinal data analysis. Our methods have a number of appealing features: the applicability is broad because the methods are developed for a unified framework with marginal generalized linear models; model assumptions are minimal in that no full distribution is required for the response process and the distribution of the mismeasured covariates is left unspecified; and the implementation is straightforward. To justify the proposed methods, we provide both theoretical properties and numerical assessments.

EMAIL: yyi@uwaterloo.ca

30. Modeling and Analysis of High Dimensional Signals and Data Applications to Imaging

› Genetics, Brain Signals and Behavior: Integrated Exploratory Analysis

Hernando Ombao*, University of California, Irvine
Yuxiao Wang, University of California, Irvine

We present exploratory approaches to modeling dependence between components of high-dimensional time series and to study its association to genetics and behavior in a decision making tasks. To address the major hurdle of high dimensionality in electroencephalograms (EEGs), we first derive factors

or summaries of activity in each brain region. Connectivity between regions is characterized through these region-specific summaries. These factor signals are computed by filtering the original EEG signals where the filter coefficients are obtained from principal components analysis of the spectral matrix. Dependence between brain regions are then explored through the regionally-derived summary factors. Spectral measures are utilized to identify frequency-specific oscillations that drive activity between pairs of regions. These exploratory approaches will be illustrated on resting-state EEG data which serves as a prelude to identifying connectivity structures that are predictive of learning. We shall also explore the role of genetics in the association between physiological signals and behavior.

EMAIL: hombao@uci.edu

› Interpretable High-Dimensional Inference Via Score Maximization in Neuroimaging

Simon N. Vandekar, University of Pennsylvania
Philip T. Reiss, University of Haifa and New York University
Russell T. Shinohara*, University of Pennsylvania

In neuroimaging a key goal is testing the association of a single outcome with a very high-dimensional imaging or genetic variable. Oftentimes summary measures of the high-dimensional variable are created to sequentially test and localize the association with the outcome. In some cases, the results for summary measures are significant, but subsequent tests used to localize differences are underpowered and do not identify regions associated with the outcome. We propose a generalization of Rao's score test based on maximizing the score statistic in a linear subspace of the parameter space. If the test rejects the null, then we provide methods to localize signal in the high-dimensional space by projecting the scores to the subspace where the score test was performed. This allows for inference in the high-dimensional space to be performed on the same degrees of freedom as the score test, effectively reducing the number of comparisons. We illustrate the method using the Alzheimer's Disease Neuroimaging Initiative dataset. Simulation results demonstrate the test has competitive power relative to others commonly used. Joint work with Simon Vandekar and Phil Reiss.

EMAIL: rshi@upenn.edu

› Learning Neuronal Functional Connectivity from Spike Train Data

Shizhe Chen, Columbia University
Daniela Witten, University of Washington
Ali Shojaie*, University of Washington

We focus on the task of learning the functional connectivity among neurons from spike trains obtained from calcium imaging data. Spike trains are sequences of timestamps for when the neurons fire. We model the spike trains as a multivariate Hawkes process, and used this framework to infer the functional connectivity among neurons. Specifically, we define the functional connectivity to be a directed graph among neurons, where a directed edge indicates that a spike in one neuron change the firing rate of the other neuron. Using the framework of multivariate Hawkes processes, we develop efficient methods for learning the edges of the directed functional connectivity network. We investigate the asymptotic properties of the proposed estimator and illustrate it using simulated and real data.

EMAIL: ashojaie@uw.edu

› Testing for SNP-Brain Network Associations

Wei Pan*, University of Minnesota

There has been increasing interest in developing more powerful and flexible statistical tests to detect genetic associations with multiple traits, as arising from neuroimaging genetic studies. We develop powerful statistical methods to estimate a brain functional network and its modules, then test their associations with genetic variants. The promising performance of the proposed methods was demonstrated with applications to the Alzheimer's Disease Neuroimaging Initiative (ADNI) data, in which either structural MRI-based phenotypes or resting-state functional MRI (rs-fMRI) derived brain functional connectivity were used as multiple phenotypes.

EMAIL: weip@biostat.umn.edu

31. Recent Developments in Optimal Treatment Regimes for Precision Medicine

How Can Psychiatric Research be SMART?

Yu Cheng*, University of Pittsburgh

A sequential multiple assignment randomized trial (SMART) is designed to identify optimal intervention sequences. Recently it has drawn great attention in psychiatry, since psychiatric treatments are often long-term and dynamic, and psychiatrists routinely tailor treatments based on their subject knowledge. However, systematic evaluations of optimal treatment strategies are generally lacking. A SMART design is uniquely suited to address questions about when to deliver which intervention to treat patients and achieve optimal long-term outcomes. We will discuss two SMART applications to Psychiatric research. One study proposes a SMART to test combinations of prenatal and postpartum interventions, where gestational weight gain is a natural tailoring variable. The other uses a non-restrictive SMART to test the efficacy of adding varenicline, exercise, or both to help quit smoking and maintain long-term abstinence. Unique features will be discussed in each application. These SMART studies will generate much-needed data on treatment sequences, which will provide empirical evidence on optimal treatment strategies and greatly improve patient wellbeing.

EMAIL: yucheng@pitt.edu

List-Based Treatment Regimes

Yichi Zhang, Harvard University
Eric B. Laber*; North Carolina State University
Marie Davidian, North Carolina State University
Butch Tsiatis, North Carolina State University

Precision medicine is currently a topic of great interest in clinical and intervention science. We formalize precision medicine as a sequence of decision rules, on per stage of clinical intervention, that map up-to-date patient information to a recommended treatment. It is well-known that that even under simple generative models the optimal treatment regime can be a highly nonlinear function of patient information. Consequently, recent methodological research has focused on the development of flexible models for estimating optimal treatment regimes. However, in many settings, estimation of an optimal treatment regime is an exploratory analysis intended to generate new hypotheses for subsequent research and not to directly dictate treatment to new patients. In such settings, an estimated regime that is interpretable in a domain context may be of greater value than an unintelligible "black-box." We propose an estimator of an optimal treatment regime composed of a sequence of decision rules, each expressible as a list of "if-then" statements that can be presented as either a paragraph or a flowchart which is immediately interpretable to domain experts.

EMAIL: laber@stat.ncsu.edu

Some Recent Developments in Machine Learning and Precision Medicine

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

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

EMAIL: kosorok@unc.edu

A Case Study in Precision Medicine: Rilpivirine Versus Efavirenz for Treatment-Naive HIV Patients

Zhiwei Zhang*, University of California, Riverside
Wei Liu, Harbin Institute of Technology
Lei Nie, U.S. Food and Drug Administration
Guoxing Soon, U.S. Food and Drug Administration

Rilpivirine and efavirenz are two drugs for treatment-naive adult patients infected with human immunodeficiency virus (HIV). Two randomized clinical trials comparing the two drugs

suggested that their relative efficacy may depend on baseline viral load and CD4 cell count. Here we estimate individualized treatment regimes that attempt to maximize the virologic response rate or the median of a composite outcome that combines virologic response with change in CD4 cell count (dCD4). To estimate the target quantities for a given treatment regime, we use G-computation, inverse probability weighting (IPW) and augmented IPW methods to deal with censoring and missing data under a monotone coarsening framework. The resulting estimates form the basis for optimization in a class of candidate regimes indexed by a smaller number of parameters. A cross-validation procedure is used to remove the re-substitution bias in evaluating an optimized treatment regime.

EMAIL: zhiwei.zhang@ucr.edu

32. Basket Trials in Oncology: Novel Designs for Targeted Treatments

› A Bayesian Design for Basket Clinical Trials

Richard Macey Simon*, National Cancer Institute, National Institutes of Health

The basket trial represents an early phase II discovery trial in which patients with a defined genomic alteration but multiple histologic types of tumors are selected to discover in which histologic types of tumors the targeted drug is active. In some cases the patients include those with a variety of seemingly similar genomic alterations and the trial is performed to determine which alterations sensitize the tumor to the drug. Basket trials are discovery trials rather than hypothesis testing trials; promising results of drug activity for a subset should be confirmed in a more focused follow-up trial. Most basket trials have been designed either ignoring the histologic strata of the patients or as independent trials for each such stratum. I will describe a new design for planning, monitoring and analyzing basket trials which weighs interim evidence that the histologic strata or uniform in sensitivity to the drug or behave as unrelated populations. A website for using the new design is available at https://brpnci.shinyapps.io/BasketTrials/R Simon, et al. Seminars in Oncology 43:13-18, 2016.

EMAIL: rsimon@mail.nih.gov

› Efficient Basket Trial Designs: A Two-Class Hierarchical Approach

Kristen M. Cunanan*, Memorial Sloan Kettering Cancer Center
Alexia Iasonos, Memorial Sloan Kettering Cancer Center
Ronglai Shen, Memorial Sloan Kettering Cancer Center
Colin B. Begg, Memorial Sloan Kettering Cancer Center
Mithat Gönen, Memorial Sloan Kettering Cancer Center

In previous work on phase II basket trials, we explore an approach that capitalizes on an interim heterogeneity assessment to potentially aggregate baskets. This adaptation displays substantial efficiencies in sample size and gains in power when the drug truly works in all or most baskets with modest losses when the drug works in only a single basket, when compared to independent Simon two-stage designs for each basket. Based on these results, we pursue further efficiencies and gains by implementing a more complex modeling approach, such as a two-class hierarchical model. In this approach, we classify baskets into two-classes, such as inactive versus active, based on their response rates. Within a class, we borrow information across baskets to improve estimation of basket-specific response rates, while still allowing individual baskets to stop due to futility. Using simulations, we compare three approaches: an adaptive design using a two-class hierarchical model, an adaptive design using a simple hierarchical model, and an adaptive design using independent Simon two-stage designs to evaluate the strengths and weaknesses of each approach.

EMAIL: cunanank@mskcc.org

› Adaptive Design of a Confirmatory Basket Trial

Cong Chen*, Merck & Co., Inc.

A basket design attempts to study patients with a common biomarker signature across multiple histologies. This study design has previously been used to explore experimental therapies with potentially transformative effects. We will present a general design of a Phase 3 basket trial broadly applicable to any effective therapy. Given the difficulty in indication selection, the basic idea is to prune the inactive indications at an interim analysis and pool the active indications in the final analysis. This presentation will provide

statistical details on Type I error control as well as correction for estimation bias for a general confirmatory basket design when different endpoints are used for pruning and pooling.

EMAIL: cong_chen@merck.com

› Implications of Different Methods of Borrowing Information in Basket Trials

Kert Viele*, Berry Consultants
Michelle Detry, Berry Consultants
Liz Krachey, Berry Consultants

Basket trials are becoming extremely popular due to their ability to investigate a variety of histologic subgroups within a single trial. We consider Bayesian Borrowing of information, which models the individual baskets using a hierarchical model. The priors in the hierarchical model have an effect on the operating characteristics of the trial. In situations where there is a general trend, hierarchical modeling produces higher power and lower type I error. In mixed situations, power reductions and inflated type I error are possible. We will quantify these advantages and disadvantages for a variety of priors and suggest default choices that produce good operating characteristics. To address the potential disadvantages of hierarchical models while keeping the advantages, cluster based generalizations (via Dirichlet Process or other mixtures) have also been proposed. We will illustrate the operating characteristics and interpretability of cluster based methods, and again focus on default choices of priors with reasonable operating characteristics.

EMAIL: kert@berryconsultants.net

33. ORAL POSTERS: Advances in Methods for Genetics and Genomics

33a. INVITED ORAL POSTER:

Integrative Genomic Analyses

Mahlet G. Tadesse*, Georgetown University
Marie Denis, CIRAD, France

Advances in high-throughput technologies have led to the acquisition of various types of -omic data on the same biological samples. Each data type provides a snapshot of the molecular processes involved in a particular phenotype. While studies focused on one type of -omic data have led to significant results, an integrative -omic analysis can provide a better understanding of the complex biological mechanisms involved in the etiology or progression of a disease by combining the complementary information from each data type. We investigated flexible modeling approaches under different biological relationship scenarios between the various data sources and evaluated their effects on a clinical outcome using data from the Cancer Genome Atlas project. The integrative models led to improved model fit and predictive performance. However, a systematic integration that allows for all possible links between biological features is not necessarily the best approach.

EMAIL: mgt26@georgetown.edu

33b. INVITED ORAL POSTER:

Simplified Power Calculations for Aggregate-Level Association Tests Provide Insights and Challenges for Rare Variant Association Studies

Nilanjan Chatterjee*, Johns Hopkins University
Andryi Derkach, National Cancer Institute, National Institutes of Health
Haoyu Zhang, Johns Hopkins University

Genome-wide association studies are now shifting focus from analysis of common to uncommon and rare variants with an anticipation to explain additional heritability of complex traits. As power for association testing for individual rare variants may often be low, various aggregate level association tests have been proposed to detect genetic loci that may contain clusters of susceptibility variants. Typically, power calculations for such tests require specification of large number of parameters, including effect sizes and allele frequencies of individual markers, making them difficult to use in practice. In this report, we approximate power to varying degree of accuracy using a smaller number of key parameters, including the total genetic variance explained by a given locus. Using the simplified power calculation methods, we then develop an analytic framework to obtain bounds on genetic architecture of an underlying trait given results from a genome-wide study and observe important

implications for lack or limited number of findings in many currently reported studies. A shiny application in R implementing the methods is made publicly available.

EMAIL: nilanjan10c@gmail.com

33c. pETM: A Penalized Exponential Tilt Model for Analysis of Correlated High-Dimensional DNA Methylation Data

Hokeun Sun*, Pusan National University
Ya Wang, Columbia University
Yong Chen, University of Pennsylvania
Shuang Wang, Columbia University

DNA methylation plays an important role in many biological processes and cancer progression. We previously developed a network-based penalized logistic regression for correlated methylation data, but only focusing on mean signals. We have also developed a generalized exponential tilt model that captures both mean and variance signals but only examining one CpG site at a time. In this article, we proposed a penalized Exponential Tilt Model (pETM) using network-based regularization that captures both mean and variance signals in DNA methylation data and takes into account the correlations among nearby CpG sites. By combining the strength of the two models we previously developed, we demonstrated the superior power and better performance of the pETM method through simulations and the applications to the 450K DNA methylation array data of the four breast invasive carcinoma cancer subtypes from The Cancer Genome Atlas (TCGA) project. The developed pETM method identifies many cancer-related methylation loci that were missed by our previously developed method that considers correlations among nearby methylation loci but not variance signals.

EMAIL: hsun@pusan.ac.kr

33d. Copy Number Variants Kernel Association Test with Application to Autism Studies

Xiang Zhan*, Fred Hutchinson Cancer Research Center

Copy number variants (CNVs) have been implicated in a variety of neurodevelopmental disorders, including autism spectrum disorders, intellectual disability and schizophrenia. Despite the availability of an unprecedented wealth of CNV data, methods for testing association between CNVs and disease-related traits are still under-developed due to the low prevalence and complicated multi-scale features of CNVs. We propose a novel CNV kernel association test (CKAT) in this paper. To address the low prevalence, CNVs are first grouped into CNV regions (CNVR). Then, taking into account the multi-scale features of CNVs, we first design a single-CNV kernel which summarizes the similarity between two CNVs, and next aggregate the single-CNV kernel to a CNVR kernel which summarizes the similarity between two CNVRs. Finally, association between CNVR and disease-related traits is assessed via a score test for variance components in a random effect model. CKAT is illustrated both via simulation studies and application to a real dataset examining the association between CNV and autism spectrum disorders.

EMAIL: xiangzhan9@gmail.com

33e. Genetic Association Analysis of Binary Traits in Structured Samples

Joelle Mbatchou*, University of Chicago
Mary Sara McPeek, University of Chicago

Case-control studies are commonly used to infer the genetic basis of a binary trait. Such studies can involve a structured sample, due to the presence of population structure or the inclusion of related individuals, or both, which can result in confounding if not properly accounted for during modeling. A logistic mixed model is a natural choice given the structure of the data. Yet, in practice, fitting such a model typically requires a trade-off between speed and accuracy, and reliance on the model for inference can lead to compromised type 1 error in the presence of model misspecification. We use a MCMC-based approach to fit a logistic mixed model, and we perform a retrospective test between the trait and each genotype, where the genotype is considered random and the analysis is done conditional on the trait and covariates. This approach has the advantage of robustness to various sources of model misspecification, such as the exclusion of important covariates or the presence of ascertainment (in which individuals are sampled based on trait/covariate status). Through

simulation studies and data analysis, we demonstrate the improvement of our method over previous approaches.

EMAIL: mcpeek@uchicago.edu

33f. POST: An Integrated Association Analysis in High Dimensional Genetic Data

Xueyuan Cao*, St. Jude Children's Research Hospital
E. Olusegun George, University of Memphis
Mingjuan Wang, St. Jude Children's Research Hospital
Dale Bowman, University of Memphis
Cheng Cheng, St. Jude Children's Research Hospital
Jeffery Rubnitz, St. Jude Children's Research Hospital
James Downing, St. Jude Children's Research Hospital
Stanley Pounds, St. Jude Children's Research Hospital

With high throughput technologies, investigators can address biological questions in whole genome level. In addition to gene level testing, set-based association study is getting attractive. Projection onto the Orthogonal Space Testing (POST) is proposed as a general and flexible procedure to perform association test for gene profiling data in a set level. The probe level signals of a set are first projected to an orthogonal subspace which is spanned by first handful eigenvectors explaining a predefined fraction of variation of probes. The projected data are subjected to parametric association tests. A POST statistic is defined as sum squares of individual z-statistic weighted by corresponding eigenvalues (a quadratic form). The p-value is approximated by a generalized χ^2 distribution. POST was applied to a gene profiling data set of 105 pediatric AML patients. Out of the 875 biological processes, 17 were significantly associated with event-free survival at q value of 0.4, 5 of which were signaling pathways including 'Regulation of Wnt receptor signaling pathway'. This indicates that POST can identify meaningful gene sets with clinical phenotypes.

EMAIL: xueyuan.cao@stjude.org

33g. Correcting for Population Stratification Induced by Spatial Variation in Phenotypic Mean in Sequencing Studies

Yeting Du*, Harvard University
Han Chen, Harvard University
Xihong Lin, Harvard University

Population stratification is a major confounding factor in sequencing studies and can lead to spurious associations. It has recently been shown that rare variants tend to exhibit a stronger stratification than common variants, especially when the phenotype has a sharp spatial distribution. The resulting inflation in test statistics may not be corrected for by popular methods such as principal components (PCs) adjustment and linear mixed models (LMMs). In this talk, we propose to account for the spatial variation in the mean of the phenotype using a thin-plate smoothing spline based on the first two PCs. We show that the smoother can be embedded in an LMM and develop SNP-set tests such as the sequence kernel association test and burden test for this model. We demonstrate via simulations that our method effectively controls for stratification in spatially structured populations when the phenotype has a sharp spatial distribution.

EMAIL: yed550@mail.harvard.edu

33h. Inverse Normal Transformation for Genome Wide Association Testing of Quantitative Traits

Zachary R. McCaw*, Harvard University
Xihong Lin, Harvard University

Genome wide association studies (GWAS) aim to identify genetic variants associated with complex diseases and phenotypic traits. Association testing for quantitative traits by parametric methods assumes normally distributed residuals. When the residual distribution is markedly non-normal, practitioners either resort to non-parametric or asymptotic tests, or they transform the outcome to achieve residual normality. Although rank-based inverse normal transformation (INT) is often used for this purpose, it currently lacks theoretical or empirical justification. In particular, INT of the outcome in a regression model does not ensure residual normality. Here we study the performance of association tests that utilize the INT to identify conditions under which testing is

valid. Through simulation, we show that methods using INT can provide valid inference in situations where simple linear regression fails; namely, when testing against non-normal phenotypes in the presence of linkage disequilibrium. To obviate the need for choosing among different INT-based methods, we combine the valid approaches into an omnibus test, which is well powered against a range of non-normal phenotypes.

EMAIL: zmccaw@g.harvard.edu

33i. Adaptive Projection Global Testing for High Dimensional Data with Dependent Structure

Jingwen Zhang*, University of North Carolina, Chapel Hill
Joseph Ibrahim, University of North Carolina, Chapel Hill
Qiang Sun, Yale University
Hongtu Zhu, University of North Carolina, Chapel Hill

The aim of this paper is to develop an Adaptive Projection Global Testing (APGT) procedure to perform hypothesis testing of a set of covariates in multivariate regression modeling with a large number of responses. We propose a dimension reduction strategy by taking advantage of correlations among multivariate responses. A fast and efficient screening procedure base on marginal statistics is first performed to select candidate signal set. Then projection transformation is adopted to maximize asymptotic signal-to-noise ratio. Numerical simulations have shown that APGT outperforms many other state-of-the-art methods when dealing with dependent data structure. In real data example, APGT is applied to ADNI data to explore important genetic markers associated with brain structure affected by Alzheimer's disease.

EMAIL: jingwn.zhang@gmail.com

33j. An Adaptive Test on High Dimensional Parameters in Generalized Linear Models

Chong Wu*, University of Minnesota
Gongjun Xu, University of Minnesota
Wei Pan, University of Minnesota

Several tests for high dimensional generalized linear models have been proposed recently, however, they are mainly based on a sum of squares of the score vector and only powerful under certain limited alternative hypotheses. In practice, since the signals in a true alternative hypothesis may be sparse or dense or between, the existing tests may or may not be powerful. In this paper, we propose an adaptive test that maintains high power across a wide range of scenarios. To calculate its p-value, its asymptotic null distribution is derived. We conduct simulations to demonstrate the superior performance of the proposed test. Then we apply it and other existing tests to an Alzheimer's Disease Neuroimaging Initiative data set, detecting possible associations between Alzheimer's disease and sets of a large number of single nucleotide polymorphisms. As an end product, we put R package GLMaSPU implementing the proposed test on GitHub and CRAN.

EMAIL: wuxx0845@umn.edu

33k. A Distributed Analysis Method for Detecting Genetic Interactions for Complex Diseases in Large Research Consortia

Yulun Liu*, University of Pennsylvania
Elisabetta Manduchi, University of Pennsylvania
Jason Moore, University of Pennsylvania
Paul Scheet, University of Texas
MD Anderson Cancer Center
Yong Chen, University of Pennsylvania

Due to small effect sizes and stringent requirements for multiple testing correction, identifying gene-gene interactions in complex diseases is more challenging than the analysis of genetic main effects alone in small independent research groups. To address the above challenges, many genomics research groups and the related initiatives collaborate to form large-scale consortia and develop open access to enable wide-scale sharing of genome-wide association study (GWAS) data. Despite the perceived benefits of data sharing from large consortia, some potential issues are raised by data sharing. In this paper, we develop a novel two-stage testing procedure, named as phylogenY-based Effect-size Tests for Interactions using first 2 moments (YETI-2), to detect gene-gene interactions through both pooled marginal effects and heterogeneity across study sites using a meta-analytic framework. This proposed method can not only be applied to distributed GWAS databases without shared individual

patient-level information but also can be used to leverage site-specific heterogeneity among sites in which the same phenotypes were measured.

EMAIL: yulunliu@mail.med.upenn.edu

33I. Robust Tests for Additive Gene-Environment Interaction in Case-Control Studies Using Gene-Environment Independence

Gang Liu*, Harvard University
Bhramar Mukherjee, University of Michigan

There have been proposals advocating the use of additive gene-environment interaction as a more relevant measure for public health actions and interventions. Several authors have characterized the gain in efficiency for estimating the multiplicative gene-environment interaction parameter in case-control studies by exploiting the assumption of gene-environment independence using the retrospective likelihood. In this paper, we propose a robust test for additive interaction in case-control studies that adaptively uses the gene-environment independence assumption and provides superior power compared to the standard test based on logistic regression. We use data from the Ovarian Cancer Association Consortium to illustrate the methods.

EMAIL: gang_liu@g.harvard.edu

34. Environmental and Geographic Applications

› Generating Partially Synthetic Geocoded Public Use Data with Decreased Disclosure Risk Using Differential Smoothing

Harrison Quick*, Drexel University
Scott H. Holan, University of Missouri
Christopher K. Wikle, University of Missouri

When collecting geocoded confidential data with the intent to disseminate, agencies often resort to altering the geographies prior to making data publicly available. An alternative to releasing aggregated and/or perturbed data is to release synthetic data, where sensitive values are replaced with draws from models designed to capture distributional features in the collected data. The issues associated with spatially outlying observations in the data, however, have received relatively little attention. Our goal here is to shed light on this problem, propose a solution -- referred to as "differential smoothing" -- and illustrate our approach using sale prices of homes in San Francisco.

EMAIL: harryq@gmail.com

› Long-Term Exposure to Multiple Pollutants and Asthma Prevalence

Joshua P. Keller*, Johns Hopkins Bloomberg School of Public Health
Roger D. Peng, Johns Hopkins Bloomberg School of Public Health

While short-term exposure to high levels of air pollution have been shown to be associated with acute asthma exacerbations, less is known about the relationship between long-term air pollution exposure and asthma. Medicaid records of children and youth in the United States provide a large, geographically diverse cohort to investigate this relationship. However, traditional analyses of this type of administrative data are limited to counties containing pollution monitors. To greatly expand the size of the cohort, we use spatial models to predict long-term exposures in all counties of the contiguous United States. We compare two approaches to incorporating multi-dimensional pollution measurements: joint modeling of all pollutants and clustering followed by prediction. These methods provide an approach to investigating variations in the pollution-asthma relationship by pollution composition.

EMAIL: jkelle46@jhu.edu

Multivariate Left-Censored Bayesian Model for Predicting Exposure Using Multiple Chemical Predictors During the Deepwater Horizon Oil Spill Clean-up

Caroline Groth*, University of Minnesota
Sudipto Banerjee, University of California, Los Angeles
Gurumurthy Ramachandran, Johns Hopkins University
Mark R. Stenzel, Exposure Assessments Applications, LLC
Dale P. Sandler, National Institute of Environmental Health Sciences, National Institutes of Health
Aaron Blair, National Cancer Institute, National Institutes of Health
Lawrence S. Engel, University of North Carolina, Chapel Hill
Richard K. Kwok, National Institute of Environmental Health Sciences, National Institutes of Health
Patricia A. Stewart, Stewart Exposure Assessments, LLC

In April 2010, the Deepwater Horizon oil rig caught fire and sank, sending approximately 5 million barrels of oil into the Gulf of Mexico over the ensuing 3 months. Thousands of workers were involved in the response and clean-up efforts. Many harmful chemicals were released into the air from crude oil including total hydrocarbons, benzene, toluene, ethylbenzene, xylene, and hexane. NIEHS's GuLF STUDY investigators are estimating the exposures the workers experienced related to the event and evaluating associations between the exposures and detrimental health outcomes. A high percentage of the measurements were below the analytical methods' limits of detection, i.e. censored. We describe a method for developing estimates of exposure when multiple sources contribute to an exposure. We model our response Y as dependent on multiple chemical predictors X while allowing censoring to be present in any combination of the Xs and Y. Using this multivariate framework accounting for censoring, we are able to develop stronger, unbiased exposure estimates, including marginal and conditional means and variance estimates for different groups of workers.

EMAIL: groth203@umn.edu

Discovering Structure in Multiple Outcomes Models for Environmental Exposure Effects

Amy A. LaLonde*, University of Rochester
Tanzy Love, 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 effect of prenatal methylmercury exposure on childhood neurodevelopment. Rather than cluster individual subjects, we cluster neurodevelopmental test outcomes within a multiple outcomes model to improve estimation of the exposure effect and the model fit diagnostics. By using information in the data to nest the outcomes into domains, the model more accurately reflects the shared characteristics the domains of neurodevelopment intend to represent. The Bayesian paradigm allows for sampling from the posterior distribution of the grouping parameters; thus, inference can be made on the groups and their defining characteristics. We avoid the often difficult and highly subjective requirement of a priori identification of the total number of groups by incorporating a Dirichlet process prior. In doing so, we allow the data to inform the selection of the appropriate number of groups as well as the group arrangement.

EMAIL: amy_lalonde@urmc.rochester.edu

Lagged Kernel Machine Regression for Identifying Time Windows of Susceptibility to Exposures of Complex Metal Mixtures

Shelley Han Liu*#, Harvard University

Exposures to metal mixtures during early life may impact cognitive function, and there may exist critical time intervals during which vulnerability is increased. However, there is a lack of statistical methods to study time-varying exposures of complex toxicant mixtures. Therefore, we develop a flexible statistical method, Lagged Kernel Machine Regression (LKMR), to identify critical exposure windows of chemical mixtures that accounts for complex non-linear and non-additive effects of the mixture at any given exposure window. LKMR is a Bayesian hierarchical model that estimates how the effects of mixture exposures change with the exposure window using a novel grouped, fused Lasso for Bayesian shrinkage. Simulation studies demonstrate the performance of LKMR under realistic exposure-response scenarios, and demonstrate large gains over approaches that consider each

critical window separately, particularly when serial correlation among the time-varying exposures is high. We apply LKMR to analyze associations between neurodevelopment and metal mixtures in PROGRESS, a prospective cohort study on metal mixture exposures and neurodevelopment conducted in Mexico City.

EMAIL: shl159@mail.harvard.edu

› Data Fusion Techniques for Estimating the Relative Abundance of Rare Species

Purna S. Gamage*, Texas Tech University
Souparno Ghosh, Texas Tech University

Estimating relative abundance of species is one of the most important problems arising in ecology. Traditionally such estimates are obtained using capture-mark-recapture methodologies. Non-invasive procedures, as, camera trap surveys have also been used extensively. Yet, such methodologies are not efficient when the focal species is relatively rare and exhibits cryptic behavior. Over the past decade, scent detection dogs were trained to identify the scats of focal species to assess occurrence of that species in a particular geographical region. Besides detection of presence, the relative abundance could also be estimated from DNA analyses of the collected scats. In this study we develop a data fusion technique to combine camera trap and scat surveys to draw inference on the relative abundance of the target species. The major challenge is developing a coherent model that can handle the discrete sampling protocol induced by camera traps and the continuous search paths of scat surveys. We extend the spatial capture-recapture model (Chandler & Royle,2013) to combine these two types of data and illustrate its application on a swift fox study conducted in north-western Texas.

EMAIL: purna.s.gamage@ttu.edu

35. Methods for HIV and Infectious Disease Modeling Research

› State Space Models of Retention, Disengagement, and Re-Entry into HIV Care

Hana Lee, Brown University
Xiaotian Wu*, Brown University
Michael J. Mugavero, University of Alabama, Birmingham
Stephen R. Cole, University of North Carolina, Chapel Hill
Bryan Lau, Johns Hopkins Bloomberg School of Public Health
Joseph W. Hogan, Brown University

Retention in care is essential to optimizing HIV patient outcomes. However, few studies examined longitudinal patient dynamics related to retention in care. We used a new approach, the state space models (SSM), to describe longitudinal patterns of engagement in care. Based on four states: engaged in care, disengaged from care, lost from care (LFC), and death, we examined retention dynamics and identified sub-groups at higher risk of falling out of care via SSM. Our data from the CFAR Network includes 31,376 patients with 20 years of follow-up information. Among engaged patients, probabilities of retention, disengagement and death are 86%, 13%, and 1%. Among disengaged, probabilities of continued disengagement, lost, and death are 24%, 58%, 16%, and 2%. The SSM identified that patients with lower CD4 counts, higher viral load, and no ARV have lower retention rate. In addition, age, gender, sexual orientation and race/ethnicity are associated with disengagement and LFC. SSM provides a regression framework for modeling longitudinal patterns of engagement in care and can be used to generate predicted outcomes for specific patient profiles and enhance retention rate.

EMAIL: xiaotian_wu@brown.edu

› Analysis of Infectious Disease Transmission Data: Binomial Considered Harmful

Yushuf Sharker*, University of Florida
Eben Kenah, University of Florida

One of the primary goals of household surveillance studies of infectious disease is to calculate the secondary attack

rate (SAR), the probability of disease transmission from an infected household member A to a susceptible member B during A's infectious period. In a household of size m with a single index case, the number of secondary infections is often treated as a binomial (m-1, p) random variable where p is the SAR. This assumes that all subsequent infections in the household are transmitted directly from the index case. Because a given transmission chain of length k has probability pk, it is thought that transmission chains of length $k > 1$ can be ignored when p is small. However, there are P(m – 1, k) such chains, so the probability of k generations of infection within the household can be much greater than the probability of any single transmission chain of length k. In simulations, we show that estimation of the SAR using a binomial model is biased upward and produces confidence intervals with poor coverage probabilities. Chain binomial models or survival analysis can be used to estimate the SAR more accurately.

EMAIL: yushuf@ufl.edu

› Bayesian Modeling and Inference for Nonignorably Missing Longitudinal Binary Response Data with Applications to HIV Prevention Trials

Jing Wu*#, University of Connecticut
Joseph G. Ibrahim, University of North Carolina, Chapel Hill
Ming-Hui Chen, University of Connecticut
Elizabeth D. Schifano, University of Connecticut
Jeffrey D. Fisher, University of Connecticut

Missing data are frequently encountered in longitudinal clinical trials. To better monitor and understand the progress over time, we must handle the missing data appropriately and determine the missing data mechanism. In this article, we develop a new probit model for longitudinal binary response data. It resolves the well-known weak identifiability issue of the variance for the random effects, and substantially improves the convergence and mixing of Gibbs sampling. We show that when improper uniform priors are specified for the regression coefficients of the joint multinomial model for the missing data indicators under nonignorable missingness, the joint posterior distribution is improper. A variation of Jeffreys's prior is thus established as a remedy for the improper posterior distribution. In addition, an efficient Gibbs sampling algorithm is developed using a collapsing technique. The proposed methodology is applied to analyze real data from an HIV prevention clinical trial. A sensitivity analysis is carried out to assess the robustness of the posterior estimates under different prior specifications and missing data mechanisms.

EMAIL: jing.wu@uconn.edu

› A Comparison of the Test-Negative and the Ordinary Case-Control Designs for Estimation of Influenza Vaccine Effectiveness under Nonrandom Vaccination

Meng Shi*, Emory University
Qian An, Emory University
Kylie Ainslie, Emory University
Michael Haber, Emory University

Observational studies are being increasingly used to estimate vaccine effectiveness (VE). We developed a probability model for comparing the bias and the precision of VE estimates from the ordinary case-control (OCC) design and the test-negative (TN) design. Our model includes the following parameters: the probability of becoming vaccinated against influenza, the probabilities of developing influenza and non-influenza acute respiratory illnesses (ARIs), and the probabilities of seeking medical care. These probabilities may depend on the subject's health status. Two outcomes of interest are considered: symptomatic influenza (SI) and medically-attended influenza (MAI). Our results suggest that if vaccination does not affect the probability of non-influenza ARI, then VE estimates from TN studies usually have smaller bias. Interestingly, when the outcome of interest is SI, the bias resulting from differences between vaccinees and non-vaccinees with respect to their health status is smaller than the other sources of bias. In summary, the TN design produces valid estimates of under certain conditions.

EMAIL: meng.shi@emory.edu

› Real-Time Epidemiological Search Strategy

Erica Billig*, University of Pennsylvania
Jason A. Roy, University of Pennsylvania
Michael Z. Levy, University of Pennsylvania

Controlling outbreaks of vector-borne diseases is often focused on containing the vector itself. We propose a novel stochastic

compartmental model to analyze urban insect infestations in the context of the re-emerging Chagas disease vector, Triatoma infestans, in Arequipa, Peru. Our approach incorporates both the counts of disease vectors at each observed house and the complex spatial dispersal dynamics. Our goal of the analysis is to predict and identify houses that are infested with T. infestans for entomological inspection and insecticide treatment. A Bayesian method is used to augment the observed data, estimate the insect population growth and dispersal parameters, and determine posterior infestation risk probabilities of households. We investigate the properties of the model with simulation studies and analyze the Chagas disease vector data to create an informed control strategy. We implement the strategy in a region of Arequipa by inspecting houses with the highest posterior probabilities of infestation and report the results from the field study.

EMAIL: ebillig@mail.med.upenn.edu

› Estimating Relative Risk of HIV in India

Chandrasekaran Kandhasamy, National Institute for Research in Tuberculosis, India
Kaushik Ghosh*, University of Nevada, Las Vegas

A model-based approach for estimatng relative risk of HIV in India is presented. The proposed method uses various conditional autoregressive models to account for spatial dependence and can incorporate any available covariate information. These models are fit to the 2011 HIV data and the best fitting model is used to obtain estimates of HIV relative risk for each state in India. The resulting model can also be used to identify effective strategies for lowering the risk of HIV infections.

EMAIL: kaushik.ghosh@unlv.edu

› Statistical Models for Estimating HIV/AIDS Epidemics with Multiple Types of Prevalence Data

Ben Sheng*, The Pennsylvania State University
Kimberly Marsh, UNAIDS
Aleksandra B. Slavkovic, The Pennsylvania State University
Jeffrey W. Eaton, Imperial College London
Le Bao, The Pennsylvania State University

Objective: We describe statistical models for HIV prevalence data collected through routine testing (RT) of pregnant women attending antenatal clinics (ANCs). RT data, unlinked anonymous testing for sentinel surveillance (SS), and nationally-representative population-based surveys (NPS), are together used to estimate HIV prevalence in Spectrum. Methods: We review the existing statistical models for SS and NPS data, and propose new statistical models for using RT data at the facility level (site-level) and aggregated regionally (census-level). We use a synthetic RT dataset to measure the impact of RT data on overall prevalence trends. Results: The fits with synthetic RT data can generally recover the underlying trend and other parameters, and improve our understanding of how SS and RT data are differently biased. Conclusion: We have proposed methods to incorporate RT data into Spectrum to obtain reasonable estimates of prevalence and other measures of the epidemic. Many assumptions underlie the utilization of RT prevalence for monitoring HIV epidemic trends, which should be tested as data from RT programs increasingly becomes available in countries.

EMAIL: bxs416@psu.edu

36. Subgroup Identification and Inference

› A Comparative Study of Subgroup Identification Methods for Differential Treatment Effect: Performance Metrics and Recommendations

Yang Chen*, State University of New York at Buffalo
Demissie Alemayehu, Pfizer Inc.
Marianthi Markatou, State University of New York at Buffalo

Subgroup identification with differential treatment effects serves as an important step towards precision medicine, as it provides evidence regarding how individuals with specific characteristics respond to a given treatment. This knowledge not only supports the tailoring of treatment strategies but also prompts the development of new treatments. This manuscript provides a brief overview of the issues associated with the methodologies aimed at identifying subgroups with differential treatment effects, and studies in depth the operational characteristics of five data-driven methods that have appeared recently in the literature. The performance of the methods under study to identify correctly the covariates affecting treatment effects is evaluated via simulation

and under various conditions. Two clinical trial data sets are also used to illustrate the application of these methods. Discussion and recommendations pertaining to the use of these methods are also provided, with emphasis on the relative performance of the methods under the conditions studied.

EMAIL: ychen57@buffalo.edu

A General Framework for Subgroup Identification and Treatment Scoring

Shuai Chen*, University of Wisconsin, Madison
Lu Tian, Stanford University
Tianxi Cai, Harvard School of Public Health
Menggang Yu, University of Wisconsin, Madison

Many statistical methods have recently been developed for identifying subgroups of patients who may benefit from different treatments. Compared with traditional outcome-modeling approaches, these methods focus on modeling interactions between treatments and covariates while minimize modeling the main effects of covariates because the subgroup identification only depends on the sign of the interaction. However these methods are scattered and often narrow in scope. We propose a general framework by weighting and A-learning, for subgroup identification in clinical studies. Our framework involves minimum modeling for the relationship between outcome and covariates pertinent to the subgroup identification. We may also estimate the magnitude of the interaction, which leads to construction of scoring system measuring the individualized treatment effect. The proposed methods are quite flexible and include many recently proposed estimators. Our approaches also allow possible efficiency augmentation and regularization for high-dimensional data. We examine the empirical performance of several procedures belonging to the proposed framework through extensive numerical studies.

EMAIL: schen264@wisc.edu

Detecting Autoimmune Disease Subsets for Estimated Autoantibody Signatures

Zhenke Wu*, University of Michigan
Livia A. Casciola-Rosen, Johns Hopkins University
Antony Rosen, Johns Hopkins University
Scott L. Zeger, Johns Hopkins University

Autoimmune diseases are human immune system's responses to autoantigens in which the body produces specific autoantibodies that target these autoantigens but also cause tissue damage. The autoantibody composition is strikingly different among patients, as indicated by the many different radiolabeled patterns obtained from the mass-spectrometry-based technology - gel electrophoresis autoradiograms (GEA). Human recognition of patterns is not optimal when the patterns are composite/complex, or patterns are scattered across a large number of samples. However, multiple sources of error (including irrelevant intensity differences across gels and warping of the gels) have traditionally precluded automation of pattern discovery using autoradiograms. In this paper, we overcome these limitations by novel initial gel data preprocessing and then propose Bayesian latent class models to detect disease subgroups. The model assumes each latent class is defined by a unique autoantibody signature. The identification and estimation of prevalence of each signature are done by Gibbs sampling. We demonstrate the utility of the proposed methods with GEA data from scleroderma patients.

EMAIL: zhenkewu@umich.edu

Development of Neurocognitive Subtypes in the Clinical Antipsychotic Treatment Intervention Effectiveness (CATIE): Schizophrenia Trial

Allison C. Fialkowski*, University of Alabama, Birmingham

Neurocognition is known to be severely impaired in schizophrenic patients. Subtypes may improve individualized therapy by predicting which antipsychotic medications improve cognitive ability over time. This analysis subtyped 468 subjects using 10 baseline assessments over 5 domains: Processing Speed, Reasoning, Verbal Memory, Working Memory, and Vigilance. Multiple imputation methods were compared to replace missing clinical factors and test scores

(1-23% missing rate). Factor Mixture Analysis was used on the 35 training imputation sets to compare 2 and 3 class, 1 factor solutions. A 2 class-1 factor model using PANSS total negative score, education level, and years since 1st prescribed an antipsychotic as covariates was chosen as the best model. Good class separation was achieved (mean entropies: training 0.722; testing 0.678). Class 2 (291, 62.2%) had a higher factor mean than Class 1 (177, 37.8%), indicating higher overall neurocognitive ability. Individuals that had experienced symptoms longer, with more negative symptoms, or with fewer years of education had a higher estimated probability of being in Class 1.

EMAIL: allijazz@uab.edu

› Subgroup Inference for Multiple Treatments and Multiple Endpoints

Patrick M. Schnell*, University of Minnesota
Qi Tang, AbbVie, Inc.
Peter Mueller, University of Texas, Austin
Bradley P. Carlin, University of Minnesota

Many new experimental treatments outperform the current standard only for a subset of the population. Additionally, when more than two treatments and multiple endpoints are under consideration, there may be many possible requirements for a particular treatment to be beneficial. In this presentation we adapt the decision-theoretic notion of admissibility to the context of evaluating treatments in such trials, especially those which seek to identify which subpopulation benefits from treatment. As an explicit demonstration of admissibility concepts we combine our approach with the method of credible subgroups, which in the case of a single outcome and treatment comparison provides Bayesian bounds on the benefiting subpopulation. Our methods account for multiplicity while showing patient covariate profiles that are (or are not) likely to be associated with treatment benefit, and are thus useful in their own right or as a guide to patient enrollment in a second stage study. We investigate our methods' performance via simulation, and apply them to a recent dataset from an Alzheimer's disease treatment trial.

EMAIL: schn0956@umn.edu

37. Imputation Approaches with Missing Data

› Multiple Imputation Inference for Multilevel Data with Missing Values: The Application of Marginalized Multilevel Model

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

In multilevel problems, incompletely-observed observational units pose a serious complexity as they can be arbitrarily seen in all levels. Fully-parametric multiple imputation (MI) inference has been increasingly used as a viable solution in many settings. We also consider MI in settings where the imputation models are based on the marginalized multilevel models (MMM). Our specific motivation in working with MMM is due to their computational flexibility as well their population-average and subject-specific interpretation. Using model-fitting techniques described by Heagerty&Zeger (2000), we develop a variable-by-variable imputation technique to particularly deal with missing values in survey settings. We illustrate our techniques in a comprehensive simulation study which assesses the algorithms repetitive sampling characteristics.

EMAIL: lg.statistics@gmail.com

› Recommendations for Using Tree-Based Methods with Multiple Imputation by Chained Equations

Emily Slade, Harvard University
Melissa G. Naylor*, Pfizer Inc.

When dealing with missing data, multiple imputation by chained equations (MICE) has emerged as a promising strategy for avoiding biased estimates and invalid inferences. Within MICE, the imputation at each step can be performed using a variety of imputation methods. Past work has claimed that nonparametric tree-based imputation methods outperform parametric imputation methods in terms of bias and coverage. However, these studies fail to provide a fair comparison to parametric imputation methods since they do not follow the established recommendation that any effects in the final analysis model (including interactions) should

be included in the parametric imputation model. We use simulation to compare the performance of parametric and tree-based methods within MICE. We show that incorporating interactions into the standard parametric imputation model greatly improves performance, and we provide guidance on when tree-based methods are advantageous. We also explore previously unanswered questions regarding the specification of tree-based tuning parameters. Simulations represent diverse scenarios including a simple clinical trial and a biomarker dataset with many predictors.

EMAIL: melissa.naylor@gmail.com

› Accounting for Dependence Induced by KNN Imputation

Anvar Suyundikov, Utah State University
John R. Stevens*, Utah State University
Christopher Corcoran, Utah State University
Jennifer Herrick, University of Utah
Roger K. Wolff, University of Utah
Martha L. Slattery, University of Utah

Missing data can arise in biomedical applications for a variety of reasons, and imputation methods are frequently applied to such data. We are motivated by a colorectal cancer study where miRNA expression was measured in paired tumor-normal samples of hundreds of patients, but data for many normal samples were missing due to lack of tissue availability. We compare the precision and power performance of several imputation methods, and draw attention to the statistical dependence induced by K-Nearest Neighbors (KNN) imputation. This imputation-induced dependence has not previously been addressed by others in the literature. We demonstrate how to account for this dependence, and show through simulation how the choice to ignore or account for this dependence affects both power and type I error rate control.

EMAIL: john.r.stevens@usu.edu

› Use of Multiple Imputation and Baseline Information for Crossover Studies with Survival Endpoints

Rengyi Xu*, University of Pennsylvania
Devan Mehrotra, Merck Research Laboratories
Pamela Shaw, University of Pennsylvania

Two-period, two-treatment crossover designs are commonly used in clinical trials. These trials can include censored outcomes, such as for a novel coagulant treatment to increase clotting time or a novel sleep aid to decrease time to falling asleep. For continuous endpoints, it has been shown that baseline measurements collected before the start of each treatment period can be useful in improving the power of the analysis. Methods to achieve the same gain for survival endpoints in this setting have not been studied. In this project, we propose a method that utilizes multiple imputation and incorporates period-specific baseline observations of a failure time outcome in a crossover trial. Brittain and Follmann (2011) proposed a hierarchical rank test that focuses on the idea that preventing an event is clinically more important than delaying an event. We show through numerical studies that our proposed method can improve the efficiency relative to Brittain and Follmann's method, while additionally providing a point estimate of the treatment effect. We will also present analysis of a real data example.

EMAIL: xurengyi@mail.med.upenn.edu

› Extension of Chained Equations Imputation for Latent Ignorable Missingness

Lauren J. Beesley*, University of Michigan
Jeremy M. G. Taylor, University of Michigan

In many modeling settings, a latent variable is introduced into the modeling framework to aid in estimation, inference, or interpretation. The presence of missing covariate or outcome data presents an additional challenge, particularly when missingness depends on unmeasured information through the latent variable. We call this missingness mechanism "latent ignorable" or "latent missing at random." In this paper, we propose a generalization of the popular and simple chained equations imputation algorithm that can handle latent ignorable covariate and outcome missingness. The methods can be used in a wide range of modeling settings, and we provide several examples. We apply the proposed algorithm to study time to recurrence in patients with head and neck cancer.

EMAIL: lbeesley@umich.edu

Multiple Imputation of Accelerometer Data for Physical Activity Measurement

Michael McIsaac*, Queen's University
Lauren Paul, Queen's University

When accelerometers are not consistently worn by individuals, missing epochs of data create issues for commonly employed methods of accelerometer data analysis, and may introduce bias into physical activity measures. Multiple imputation is a potential solution to these issues, as it allows analysts to fill in the missing epochs of data with plausible values, and the usual methods analysis can subsequently be employed. However, accelerometer data possess unique challenges for multiple imputation that require careful consideration when selecting imputation models and methods, and these challenges become more daunting when imputation is performed at the epoch-level. Zero-inflated Poisson and log-normal imputation models are contrasted with simpler forms of imputation and these models are evaluated based on their epoch-level imputation accuracy as well as their ability to recover common physical activity summary measures of interest. The accelerometer data used in this investigation is from The Active Play Study, our ongoing physical activity study involving children and youth in Kingston, Ontario.

EMAIL: mikemcisaac@gmail.com

38. Experimental Design

Sample Size Re-Estimation for Confirmatory Two-Stage Multi-Arm Trials with Normal Outcomes

Yan Li*, University of Alabama, Birmingham
Jeffery M. Szychowski, University of Alabama, Birmingham

The need for efficient clinical trial designs has led to the development of flexible multi-arm designs that allow interim treatment selection and adaptation, primarily sample size re-estimation (SSR). However, SSR methods for two-arm trials are not directly applicable to multi-arm trials due to the application of the closed test procedure. In this study, we derive procedures for SSR based on the conditional power approach in the context of confirmatory two-stage multi-arm trials with normal outcomes for three designs: inverse normal combination test (INC), Fisher's combination test (FiC) and flexible group sequential design (FGS). We present extensive simulation studies to evaluate the performance of the three designs with and without SSR. Results show that, without SSR, INC and FGS are slightly more powerful than FiC. With SSR, FGS is the most powerful design, though it requires slightly higher sample sizes than INC. Meanwhile INC always outperforms FiC. In practice, the choice of the most appropriate method will depend on the actual treatment effect profile, treatment selection rule applied and the maximum number of patients allowed.

EMAIL: liyan@uab.edu

Sample Size and Power Estimations in Dynamic Risk Prediction of Cumulative Incidence Functions

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

Dynamic risk prediction has gained increasing attention as it incorporates accumulated information such as time-varying covariates and intermediate event status into risk prediction in survival analysis. Prediction is updated using landmark data set through subsetting the data with left-truncation at the landmark time and enforcing administrative censoring at the prediction horizon. Absolute difference in cumulative incidence functions (CIFs) is a widely used effect-size measurement in comparing the difference in risk for two comparison groups. In this study we first investigated the effects at different landmark and prediction horizon times on the post-hoc power of risk-difference estimation. We used the resampling method to derive the standard deviation of the estimated risk difference for both proportional hazards and non-proportional hazards settings. We then derived a formula to calculate the sample size needed to obtain the desired significance level and power in dynamic prediction. An R package was created to carry out the aforementioned estimation of difference in CIFs as well as post-hoc power and sample size calculation under various settings.

EMAIL: zhs17@pitt.edu

Sample Size Estimation Method for Evaluating the Clinical Utility of a Predictive Biomarker

Henok G. Woldu*, University of Georgia
Kevin K. Dobbin, University of Georgia

In recent years the "one drug treats all" treatment model has progressed to address an individual patient's ability to respond to a given treatment. Cancer biomarkers, mainly predictive, are currently used in clinical settings to treat individuals that are most likely to benefit and refrain treatment from the rest. Additionally, statistical methods for evaluating the clinical utility of these biomarkers have also recently advanced. In this paper, (1) we developed alternative mathematical equations for the estimator proposed by Janes et al. [2014]. This estimator, which measures the decrease in the proportion of unfavorable outcome as a result of using predictive biomarker guided treatment strategy, is used to evaluate the clinical utility of a predictive biomarker in personalized treatment. (2) We propose a sample size calculation method, Squared Mean Inverse Linear Regression, required to achieve a desired confidence mean width of the estimator. Our sample size estimation approach is based on our observation that there is a perfect linear relationship between the sample size and inverse of the confidence interval mean width squared.

EMAIL: henok535@uga.edu

Alternative Randomization Strategies for Seamless Phase I/II Adaptive Design for Trials of Novel Cancer Treatments

Donglin Yan*, University of Kentucky
Emily V. Dressler, University of Kentucky

We propose an alternative strategy for randomization on a recently proposed seamless phase I/II adaptive design. The original design by Wages and Tait was proposed for trials of molecularly targeted agents in cancer treatments, where dose-efficacy assumptions are not always monotonically increasing. We alter the calculation of dose-randomization probabilities with the goal to improve the design's performance. The proposed randomization strategy calculates randomization probabilities using the likelihood of every candidate model to as opposed to the original design that selects the best model and then randomize based on estimations from the selected model. Simulations show that under most scenarios, our revised method of randomization allocates more patients to the optimal dose while maintaining approximately the same accuracy in selecting the optimal dose without increasing the risk of toxicity. By using the revised randomization strategy, the number of patients allocated to the optimal dose is increased by approximately 10% on average, and 35% maximum. The proposed randomization strategy is most appropriate when dose-efficacy is not monotonically increasing.

EMAIL: donglin.yan@uky.edu

Comparison of Statistical Methods for the Analysis of Crossover Trials

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

The main advantage of a crossover design is to improve power as compared with a parallel study design by using each patient as his/her own control, however, crossover studies are prone to two major problems, period effect and carryover effect. The effects of including both or either of these two effects on power to detect the main treatment effects in crossover study analysis are not formally assessed. The objective of this study is to compare the performance of four methods for the analysis of crossover studies: namely (1) analysis approaches using the first period measurements only as the two sample t-test or one-way analysis of variance model; the general Gaussian linear models considering repeated measurements include (2) both period and carryover effects; (3) period effect only; (4) neither of these two effects. The type I error, power and bias are compared across the four methods through simulations for two-period and three-period crossover studies. The simulation results show that the two-period crossover design is only suggested when the carryover effect is negligible, and the three-period crossover design can be used even when substantial carryover effect exists.

EMAIL: wei.wang2@fda.hhs.gov

Stepped Wedge Cluster Randomized Controlled Trials: Sample Size and Power Determinations

Hsiang-Yu Chen, University of Pennsylvania
Jesse Y. Hsu*, University of Pennsylvania
Chung-Chou H Chang, University of Pittsburgh

Stepped wedge cluster randomized controlled trials (RCTs) are increasingly used in evaluating a causal-effect relationship between an intervention and an outcome. Sample size and power calculations are critical while designing a stepped wedge cluster RCT. The data structure of stepped wedge cluster RCTs is usually hierarchical and correlated. The mixed models approach is used to account for the correlation of observations within each level. In this presentation, we discuss linear mixed models (LMM) and generalized linear mixed models (GLMM) with the identity link function for continuous responses, and GLMM with the logit link function for binary responses. We conducted simulation studies to compute the empirical power of the hypothesis test for no intervention effect versus a pre-specified intervention effect. The purpose of this study is to evaluate the power for both the continuous and binary responses in stepped wedge cluster RCTs with three levels (e.g., hospital, physician, and individual levels). The findings provide essential information in determining the optimal sample size and also can assure adequate power for stepped wedge cluster RCTs.

EMAIL: hsu9@mail.med.upenn.edu

39. Functional Data with Applications in Biostatistics

General Additive Function-on-Function Regression

Ana-Maria Staicu*, North Carolina State University
Janet Kim, Astellas Pharma
Arnab Maity, North Carolina State University
David Ruppert, Cornell University

We consider non-linear regression models for function-on-function regression. We introduce flexible models where the mean response at a particular time point depends on the time point itself as well as the entire covariate trajectory. In this framework, we develop computationally efficient estimation methodology and discuss prediction of a new response trajectory. Additionally we discussed inference in the form of prediction intervals and hypothesis testing. The proposed estimation/inferential procedure accommodates realistic scenarios such as correlated error structure as well as sparse and/or irregular design. We investigate our methodology infinite sample size through simulations and a real data application.

EMAIL: astaicu@ncsu.edu

Statistical Framework for Joint Modeling of Rhythms of Activity, Energy, and Mood Assessed by Mobile Technologies

Vadim Zipunnikov*, Johns Hopkins Bloomberg School of Public Health
Jordan Johns, Johns Hopkins University
Junrui Di, Johns Hopkins University
Haochang Shou, University of Pennsylvania
Kathleen Merikangas, National Institute of Mental Health, National Institutes of Health

The real-time diaries available through smart-phones/watches are now extensively used for ecological momentary sampling that taps patterns of many homeostatic systems including sleep, emotional states, energy, dietary intake, and others to provide a rich set of tools to assess behavioral components of human homeostatic systems. I will, first, discuss data analytical challenges associated with modeling and interpreting data collected by electronic diaries paired with accelerometry data. I will, then, introduce a statistical framework that fully exploits the huge amount of functional/time series data that emerge from these technology. This framework provide reproducible and highly reliable measures crucial to translating diary/physical activity data into meaningful analyses of health status. Our methodology is appropriate to analyze behavioral and physiological information and can be widely used in the implementation of preventive, interventional, and translational efforts.

EMAIL: vadim.zipunnikov@gmail.com

› Partition Mixture Regression Models

Hongtu Zhu*, University of Texas MD Anderson Cancer Center
Michelle Miranda, University of Texas MD Anderson Cancer Center
Lian Heng, University of New South Wales, Australia

We propose a novel regression framework to efficiently model a scalar response variable as a function of covariates taking the form of multi-dimensional arrays, called tensors. Assume we observe data $\{(y_i, \mathcal{X}_i), i=1,\ldots,n\}$ from n subjects, where y_i is a scalar response, and $\mathcal{X}_i \in R^{p_1 \times p_2 \times \ldots \times p_D}$ is an imaging tensor covariate. Examples of $\mathcal{X}_i$ are imaging data collected over space or over space and time, e.g. functional magnetic ressonance image (fMRI), diffusion tensor image (DTI) and positron emission tomography (PET). We propose a regression framework specifically designed to address the key features of neuroimaging data: relatively low signal to noise ratio, spatially clustered effect regions, and the tensor structure of imaging data.

EMAIL: hzhu5@mdanderson.org

› Principal Component Analysis of Functional Signals over Bidimensional Manifolds, with Applications to Neuroimaging Data

Laura M. Sangalli*, Politecnico di Milano, Italy

Motivated by the analysis of high-dimensional neuroimaging signals over the cerebral cortex, we introduce a principal component analysis technique that can be used for exploring the variability and performing dimensional reduction of signals observed over two-dimensional manifolds. The proposed method is based on a PDE regularization approach, involving the Laplace-Beltrami operator associated to the manifold domain. It can be applied to data observed over any two-dimensional Riemannian manifold topology. The proposed method is applied to the study of main connectivity patterns of neural activity in the cortex, based on the analysis of a dataset made available by Human Connectome Project and consisting of resting state functional magnetic resonance imaging scans from about 500 healthy volunteers.

EMAIL: laura.sangalli@polimi.it

40. EHR Data and Bayesian Analysis: The Tail Wagging the Dog?

› Survival Analysis with Electronic Health Record Data: Experiments with Chronic Kidney Disease

Yolanda Hagar*, University of Colorado, Boulder
David Albers, Columbia University
Rimma Pivovarov, Columbia University
Herbert Chase, Columbia University
Noemie Elhadad, Columbia University
Vanja Dukic, University of Colorado, Boulder

We present a detailed survival analysis for chronic kidney disease (CKD). The analysis is based on the electronic health record (EHR) data comprising almost two decades of clinical observations collected at New York-Presbyterian, a large hospital in New York City with one of the oldest electronic health records in the United States. Our survival analysis approach centers around Bayesian multiresolution hazard modeling, with an objective to capture the changing hazard of CKD over time, adjusted for patient clinical covariates and kidney-related laboratory tests. Special attention is paid to statistical issues common to all EHR data, such as cohort definition, missing data and censoring, variable selection, and potential for joint survival and longitudinal modeling, all of which are discussed alone and within the EHR CKD context.

EMAIL: yolanda.hagar@colorado.edu

› Measuring Performance for End-of-Life Care: A Bayesian Decision-Theoretic Approach

Sebastien Haneuse*, Harvard School of Public Health
Kyu Ha Lee, The Forsyth Institute

The Centers for Medicare and Medicaid Services currently uses readmission rates to rank and profile hospitals. These rates are calculated for a range of acute health conditions for which prognosis is good, including pneumonia and heart failure. For advanced health conditions where prognosis is poor, such as pancreatic cancer, use of readmission rates alone to measure hospital performance is problematic because they ignore variation in mortality rates. Building on the semi-competing risks

framework, we propose readmission and mortality cumulative rate functions as a novel means of measuring quality of end-of-life care. Estimation of these functions follows within the Bayesian paradigm, using a recently developed hierarchical modeling framework. In addition, using a decision-theoretic approach, we also develop estimators of a range of non-standard but important inferential targets including: identification of extreme performers; ranking of hospitals; and, estimation of the empirical distribution function. The ideas and methods are illustrated using data on all Medicare beneficiaries diagnosed with pancreatic cancer between 2000-2013.

EMAIL: shaneuse@hsph.harvard.edu

› A Bayesian Latent Variables Approach to Phenotype Estimation using EHR Data

Rebecca A. Hubbard*, University of Pennsylvania

Data available in Electronic Health Records (EHR) include a wide variety of clinical assessments, biometric measurements and laboratory test results that describe patient phenotypes. However, available measurements may vary systematically across healthcare systems, clinics or patient sub-groups in ways that preclude use of standard missing data approaches. For example, depending on availability of technology, clinics may use different tests to assess the same underlying patient characteristic. This lack of overlap in measures prohibits estimation of the joint distribution needed for missing data approaches such as multiple imputation. By taking a Bayesian approach, existing evidence on the correlation between measures and the relative strength of different measures can be harnessed to facilitate phenotype estimation. In this presentation, we describe a Bayesian latent variables framework for combining the diverse data available for each individual into a summary phenotype. Through simulation studies we demonstrate the performance of this approach compared to standard missing data methods. Finally, we apply our new approach to a study of pediatric diabetes.

EMAIL: rhubb@upenn.edu

41. Estimation Of Effects in Public Health Research with Interference

› Assessing Individual and Disseminated Effects in Network HIV Treatment and Prevention Trials

Ashley L. Buchanan*, University of Rhode Island
Sten H. Vermund, Vanderbilt University School of Medicine
Samuel R. Friedman, National Development and Research Institutes, Inc.
Donna Spiegelman, Harvard School of Public Health

Implementation trials typically randomize some participants to directly receive the intervention, yet others in the network may receive the intervention through diffusion. The individual effect measures the impact on participants directly receiving the intervention above and beyond being in an intervention network. The disseminated effect measures the impact on participants sharing a network with directly-treated individuals. We show how generalized estimating equations with covariate adjustment in the outcome model can be used to account for individual-level confounding and consider two approaches for handling effect modification by index status. In the HIV Prevention Trials Network 037 Trial, there was a 28% risk reduction overall in any injection-related risk behavior (95% confidence interval (CI) = 0.59, 0.88). A 29% adjusted risk reduction in any injection-related risk behavior was observed among network members (95% CI = 0.56, 0.92). A 35% composite (individual and disseminated) risk reduction was observed in the adjusted model (95% CI = 0.47, 0.89). Methodology is now available to estimate these effects, enhancing the knowledge gained from network-randomized trials.

EMAIL: buchanan@uri.edu

› The Auto-G-Formula to Evaluate Causal Effects on a Network

Eric J. Tchetgen Tchetgen*, Harvard University
Ilya Shpitser, Johns Hopkins University

We consider the challenging goal of making inferences about causal effects on a network of connected persons. Complications are two-fold (i) the presence of Interference,

that is the possibility a person's outcome can be influenced by another's treatment; and (ii) strong outcome dependence arising from human interactions. We propose an approach that largely resolves both complications without assuming (a) partial interference, nor (b) that outcome dependence can be explained away by conditioning on network features available prior to treatment allocation. This is a notable development as several existing methods assume either (a) or (b), neither of which is realistic for most network settings. The approach is based on a recent extension of Robins' G-formula to the network setting which we call the Auto-G-Formula. In this talk, we will focus on Auto-G-Computation, a computationally efficient approach for evaluating the Auto-G-Formula for direct and spillover effects based on a single realization of a network. Both simulations and an application illustrating the methodology will be presented.

EMAIL: etchetgen@gmail.com

› Causal Inference with Interference: Estimating Total and Indirect Vaccine Effects in Cluster-Randomized Studies

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

We have shown that unbiased estimates of total, indirect, and overall effects of vaccination can be obtained using a two-stage randomized design in the presence of interference. First, clusters are randomized to a particular vaccination strategy, then individuals within the clusters are randomized according to the strategy. Many studies of vaccination randomize at the cluster level, but do not randomized at the individual level. We present several examples of such cluster-randomized studies. We discuss some of the issues related to interpreting the estimates obtained in such studies. We also consider different approaches to obtaining better estimates of total and indirect vaccine effects from such cluster-randomized designs.

EMAIL: betz@u.washington.edu

› Regression-Based Adjustment for Homophily Bias in Peer Effect Analysis

Lan Liu*, University of Minnesota
Eric J. Tchetgen Tchetgen, Harvard University

A prominent threat in social network analysis is the presence of homophily bias, that is the social influence between friends and families is entangled with common characteristics or underlying similarities that form close connections. Analysis with social network data have suggested that health related outcomes such as obesity and psychological states such as happiness and loneliness can be spread over a network. However, such analysis of peer effects or contagion effects have come under critique since homophily bias may compromise causal statement. We develop a novel approach to identify and estimate contagion effects in studies with unobserved confounding for either binary or continuous outcomes. A simulation study is carried out to investigate the finite sample performance of the proposed estimators. The methods are further illustrated in an application of evaluating the peer effect in the spread of obesity in Framingham Heart Study.

EMAIL: liu1815@gmail.com

42. Statistical Methods for Emerging High-Throughput Genomic Technologies

› Statistical Methods for Profiling Long Range Chromatin Interactions from Repetitive Regions of the Genome

Ye Zheng, University of Wisconsin, Madison
Ferhat Ay, La Jolla Institute For Allergy and Immunology
Sunduz Keles*, University of Wisconsin, Madison

Genome-wide data from chromosome conformation capture (3C-based) technologies provide overwhelming evidence that the three-dimensional (3D) organization of chromatin impacts gene regulation and genome function. Analysis of Hi-C data starts with alignment of the reads to the reference genome. In the downstream quantification step, most approaches utilize state-of-the-art bias adjustments such as proximity, GC, and mappability which are well studied in other NGS applications. However, one critical shortcoming of existing approaches is

that they discard reads that align to multiple locations on the genome, i.e., multi-reads. We study this problem in depth and train a generative model to probabilistically allocate multi-reads to their mapping locations. This highly versatile model is able to utilize auxiliary 1D epigenome data and improve allocation accuracy. Our results suggest that effective utilization of multi-reads increases sequencing depth significantly up to 15% and can, on average, identify up to 17% more enhancer-promoter interactions, which are highly reproducible across biological replicates.

EMAIL: keles@stat.wisc.edu

› Studying Intra-Tumor Heterogeneity using DNA Sequencing Data from Single Cells and Bulk-Tissue Sample

Wei Sun*, Fred Hutchinson Cancer Research Center
Chong Jin, University of North Carolina, Chapel Hill
Jonathan A. Gelfond, University of Texas Health Science Center, San Antonio
Ming-Hui Chen, University of Connecticut
Joseph G. Ibrahim, University of North Carolina, Chapel Hill

Intra-tumor heterogeneity (ITH) refers to the fact that tumor cells of one patient are heterogeneous (e.g., either spatially and temporally). For example, due to ITH, a targeted cancer drug may kill part of the tumor cells harboring the target but leave other tumor cells untouched. Precision cancer medicine aims to treat cancer patients based on their genomic lesions. A fundamental hurdle to this goal is to discern ITH based on genomic data. Severa methods have been developed to discern ITH using genomic data collected from bulk tissue samples. These methods cannot distinguish two subclones with similar cellular frequencies. Single cell sequencing (SCS) data provide extremely useful informative for ITH because if two mutations appear in a single cell, they are definitely from one subclone. We develop a new method to infer ITH using genomic data (e.g., exome-seq data) from single cells and bulk-tissue sample.

EMAIL: wsun@fredhutch.org

› Statistical Methods for Assessing Transcriptional Changes and Characterizing Heterogeneity in Single-Cell RNA-seq Data

Raphael Gottardo*, Fred Hutchinson Cancer Research Center

Single-cell RNA-seq enables the unprecedented interrogation of gene expression in single-cells. The stochastic nature of transcription is revealed in the bimodality of single-cell data, a feature shared across many single-cell platforms. I will present new methodology to analyze single-cell transcriptomic data that models this bimodality within a coherent generalized linear modeling framework. Our model permits direct inference on statistics formed by collections of genes, facilitating gene set enrichment analysis. The residuals defined by our model can be manipulated to interrogate cellular heterogeneity and gene-gene correlation across cells and conditions, providing insights into the temporal evolution of networks of co-expressed genes at the single-cell level. I will also discuss unwanted sources of variability in single-cell experiments and in particular the effect of the cellular detection rate defined as the fraction of genes turned on in a cell, and show how our model can account and adjust for such variability. Finally, I will illustrate this novel approach using several datasets that we have recently generated to characterize specific human immune cell subsets.

EMAIL: rgottardo@fredhutch.org

› Detection of Cell Types, Lineages, and Trajectories from Single-Cell Gene Expression Data

Lan Jiang, Harvard University
Huidong Chen, Harvard University
Guo-Cheng Yuan*, Harvard University

High throughput, single-cell technologies have great potential in discovering new cell types, lineages, and developmental trajectories. On the other hand, the high level of noise and low level of sensitivity in single-cell gene expression data present significant challenges for computational inference of such entities directly from data. Our group has recently developed a number of computational tools to address these challenges. Using these methods, we are able to gain new insights into developmental processes and cancer.

EMAIL: gcyuan@jimmy.harvard.edu

43. Innovative Group Sequential Methods for Biomarker Validation

Identifying Optimal Approaches to Early Termination in Two-Stage Biomarker Validation Studies

Alexander M. Kaizer, University of Minnesota
Joseph S. Koopmeiners*, University of Minnesota

Group sequential study designs (GSDs) have been proposed as an approach to conserve resources in biomarker validation studies. Typically, GSDs allow both "early termination to reject the null hypothesis" and "early termination for futility" if there is evidence against the alternative hypothesis. In contrast, several authors have advocated for using GSDs that allow only early termination for futility in biomarker validation studies due to the desire to obtain a precise estimate of marker performance at study completion. This suggests a loss function that heavily weights the precision of the estimate obtained at study completion at the expense of an increased sample size when there is strong evidence against the null hypothesis. We propose a formal approach to comparing designs by developing a loss function that incorporates the expected sample size under the null and alternative hypotheses, as well as the mean squared error of the estimate obtained at study completion. We then use our loss function to compare several candidate designs and derive optimal two-stage designs for a recently reported validation study of a novel prostate cancer biomarker.

EMAIL: koopm007@umn.edu

Two-Stage Adaptive Cutoff Design for Building and Validating a Prognostic Biomarker Signature

Mei-Yin Polley*, Mayo Clinic
Eric Polley, Mayo Clinic
Erich Huang, National Cancer Institute, National Institutes of Health
Boris Freidlin, National Cancer Institute, National Institutes of Health
Richard Simon, National Cancer Institute, National Institutes of Health

Cancer biomarkers are frequently evaluated using specimens collected from previously conducted therapeutic trials. Routine collection and banking of high quality specimens is an expensive and time-consuming process. Therefore, care should be taken to preserve these precious resources. We propose a novel two-stage adaptive cutoff design that affords the possibility to stop the biomarker study early if an evaluation of the model performance is unsatisfactory at an early stage, thereby allowing one to preserve the remaining specimens for future research. Our simulation studies demonstrate that under the null hypothesis when the model performance is deemed undesirable, the proposed design maintains type I error at the nominal level, has high probabilities of terminating the study early, and results in substantial savings in specimens. Under the alternative hypothesis, power is generally high when the total sample size and the targeted degree of improvement in prediction accuracy are reasonably large. We illustrate the use of the procedure with a dataset in patients with diffuse large-B-cell lymphoma.

EMAIL: polley.mei-yin@mayo.edu

Unbiased Estimation of Biomarker Panel Performance when Combining Training and Testing Data in a Group Sequential Design

Nabihah Tayob*, University of Texas MD Anderson Cancer Center
Kim-Anh Do, University of Texas MD Anderson Cancer Center
Ziding Feng, University of Texas MD Anderson Cancer Center

Motivated by an ongoing study to develop a screening test able to identify patients with undiagnosed Sjogren's Syndrome in a symptomatic population, we propose methodology to combine multiple biomarkers and evaluate their performance in a two-stage group sequential design as follows: biomarker data is collected from first stage samples; the biomarker panel is built and evaluated; if the panel meets pre-specified performance criteria the study continues to the second stage and the remaining samples are assayed. The design allows us to conserve valuable specimens in the case of inadequate biomarker panel performance. We propose a nonparametric conditional resampling algorithm that uses all the study data to provide unbiased estimates of the biomarker combination rule and the sensitivity of the panel corresponding to specificity of 1-t on the receiver operating characteristic curve (ROC). The Copas and Corbett (2002)

correction, for bias resulting from using the same data to derive the combination rule and estimate the ROC, was also evaluated and an improved version was incorporated.

EMAIL: ntayob@mdanderson.org

44. Measurement Error in Causal Inference

› Propensity Score-Based Estimators with Multiple Error-Prone Covariates

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

Most propensity score methods assume that the covariates are measured without error. However, covariates are often measured with error, which lead to biased causal effect estimates. Although some studies have investigated the impact of a single mismeasured covariate on estimating a causal effect and proposed methods for handling measurement error in a single covariate, almost no work exists investigating the case where multiple covariates are mismeasured. Using extensive simulation studies we examine the consequences of multiple error-prone covariates when estimating causal effects using propensity score-based estimators. We find that causal effect estimates become less biased when the propensity score model includes mismeasured covariates where the true covariates are more strongly correlated each other. However, when the measurement errors are correlated each other, additional bias is introduced. In addition, we find that when confounders are mismeasured it is beneficial to include auxiliary variables which are correlated with the true confounders in the propensity score model in terms of bias.

EMAIL: hhong@jhu.edu

› The Mechanics of Omitted Variable Bias and the Effect of Measurement Error

Yongnam Kim*, University of Wisconsin, Madison
Peter M. Steiner, University of Wisconsin, Madison

When estimating causal effects in observational studies, adjusting the effect estimates for observed covariates is the most popular way to deal with confounding bias. It is well known that the measurement error in covariates diminishes their potential to remove bias. In this paper we show that, contrary to the general belief, adjusting for an unreliably measured confounder can lead to a less biased estimate than the corresponding reliable measure. Using a simple linear regression setting with an observed confounder (X) and unobserved confounder (U), we formally investigate the effect of measurement error in X on the X-adjusted treatment effect. Measurement error in X produces a less bias if (i) X's bias-amplifying effect dominates its bias-reducing effect, or (ii) X and U induce biases in opposite directions such that their respective biases perfectly or partially offset each other. In such cases, measurement error in X attenuates bias amplification and the cancellation of offsetting biases. We discuss the practical implications of this finding, particularly for situations where we have weak subject-matter knowledge about the data-generating process.

EMAIL: ykim379@wisc.edu

› Methods to Estimate Causal Effects Adjusting for Confounding when an Ordinal Exposure is Mismeasured

Danielle Braun*, Harvard School of Public Health and Dana-Farber Cancer Institute
Marianthi-Anna Kioumourtzoglou, Columbia School of Public Health
Francesca Dominici, Harvard School of Public Health

Long-term exposure to air pollution has consistently been associated with adverse health outcomes. Most previous studies assume the exposure is error-free, but in reality the exposure is often subject to measurement error. Although some studies have addressed this issue, no study to our knowledge has done so in a causal inference framework. We propose to address this gap in the context of studying the association of long-term air pollution exposure and mortality in Medicare beneficiaries. For the entire Medicare population (main study), long-term exposure to fine particles (PM2.5) is determined from a spatiotemporal model that uses multiple different sources as input (meteorology, land use variables, satellite data, etc.). PM2.5 exposure based on these predictions is inaccu-

rate, but for a subset of zip-codes (validation study) we have actual PM2.5 concentrations measured at monitors (error-free exposure). Using this internal validation study, we develop an approach to adjust for the measurement error in the ordinal PM2.5 exposure. We apply our approach to the entire Medicare population, and estimate the association between PM2.5 exposure and mortality in this population.

EMAIL: dbraun@hsph.harvard.edu

45. ORAL POSTERS: Novel Methods for Big Data

45a. INVITED ORAL POSTER:

NeuroconductoR: Building the R Imaging Community

Ciprian Crainiceanu*, Johns Hopkins University

The number of Statisticians involved in impactful Neuroimaging research has been small. Major reasons for this bewildering reality have been the relatively steep learning curve, the lack of educational materials, and the parallel development between Neuroimaging tools and methodology on one side and R on the other side. In this talk I will describe NeuroconductoR, a platform for R packages dedicated to imaging research and inspired by the highly successful BioconductoR platform for Genomics. NeuroconductoR will use new tools that have become available including GitHuB and Travis CI. NeuroconductoR provides an organization of available imaging software in R, provides educational tools, and takes advantage of the latest versions of Neuroimaging platforms including ANTSand fsl.

EMAIL: ccraini1@jhu.edu

45b. INVITED ORAL POSTER:

Integrative Methods for Functional and Structural Connectivity

DuBois Bowman*, Columbia University
Daniel Drake, Columbia University
Ben Cassidy, 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.

EMAIL: dubois.bowman@columbia.edu

45c. Grange Mediation Analysis for Task fMRI Time Series

Yi Zhao*, Brown University
Xi Luo, Brown University

Making inference about brain effective connectivity is of great interest in task fMRI experiments. In this study, we are interested in quantifying how one brain region (mediator) intermediates the effect of external stimuli on the outcome brain region. To achieve this, we consider causal mediation analysis under structural equation modeling framework, which requires both "ignorability" and "no interference" assumptions. To address the violation of ignorability, correlation between errors is introduced to account for the unmeasured confounding effect; and to characterize the temporal and interregional dependency, Granger causality is implemented. In this paper, we propose a Granger Mediation Analysis framework that provides inference about spatial and temporal causality between brain regions using the multilevel fMRI time series. Simulation studies show that our method reduce the bias in estimating the causal effects compared to existing approaches. Applying the proposed method on a real fMRI dataset, it not only estimates the mediation effect, but effectively captures the feedback effect of the outcome region on the mediator region.

EMAIL: yi_zhao@brown.edu

45d. Bayesian Methods for Image Texture Analysis with Applications to Cancer Radiomics

Xiao Li*, University of Texas MD Anderson Cancer Center
Michele Guindani, University of California, Irvine
Chaan Ng, University of Texas MD Anderson Cancer Center
Brian Hobbs, University of Texas MD Anderson Cancer Center

Radiomics, as an emerging field in quantitative imaging, encompasses a broad class of analytical techniques. GLCM, as one type of texture features, is a matrix defined over an image to be the distribution of co-occurring gray-level pixels at a given offset and angle. GLCM-based texture features have been used to quantitatively describe tumor phenotypes. Recent literature has interrogated associations with clinical/pathology information with Machine Learning algorithms using GLCM-based texture features. Reducing the multivariate functional structure to a set of summary statistics is potentially reductive, however, masking patterns that describe disease pathogenesis. In this article, we present a Bayesian probabilistic method for count data observed over a lattice with symmetric structure. The approach was utilized to model the entire GLCM as a multivariate response surface, and applied in a cancer detection context to discriminant malignant from benign adrenal lesions using GLCMs arising from pixel-level image data. In case study and simulation study, our proposed method improved classification accuracy compared to current approaches utilizing GLCM-based texture features.

EMAIL: xiao.li.1@uth.tmc.edu

45e. Bootstrap Measures of Uncertainty in Independent Component Analysis with Application to EEG Artifact Removal

Rachel C. Nethery*, University of North Carolina, Chapel Hill
Young Truong, University of North Carolina, Chapel Hill

Independent component analysis (ICA) is a statistical method that can be applied to multivariate data generated by the linear mixing of signals in order to recover the original, unmixed signals, also known as latent sources of activity. Though ICA is a powerful tool for cleaning data collected from biomedical devices, statistical testing and inference on ICA model parameters are obstructed by challenges related to uncertainty estimation, thereby limiting its utility. In this work, we propose a bootstrap algorithm that overcomes these challenges and can be used to construct confidence intervals for ICA parameters. Through simulation studies, we demonstrate the reliable performance of the bootstrapping algorithm and provide novel insights into the estimation of source signal second moments with ICA. Finally, the impact of this work is exhibited through an application to EEG data, in which the bootstrap confidence intervals are used establish a statistical testing framework to distinguish signals of interest from noise and artifactual signals.

EMAIL: nethery@live.unc.edu

45f. Integrative Bayesian Analysis of Radiologic Imaging and Multiplatform Genomics Data

Youyi Zhang*, University of Texas MD Anderson Cancer Center
Jeffrey Morris, University of Texas MD Anderson Cancer Center
Veerabhadran Baladandayuthapani, University of Texas MD Anderson Cancer Center

We present a multi-scale Bayesian hierarchical model for integrative analysis of Radiogenomic data. Our goals are to simultaneously identify significant genes and imaging markers along with the hidden associations between these two modalities, and to further detect the overall prognostic relevance of the combined markers. For this task, we propose a multi-scale Bayesian hierarchical model which involves several innovative strategies: it incorporates integrative analysis of multi-platform genomics data sets to capture fundamental biological mechanism in a radiogenomics framework; explores the associations between imaging markers accompanying genetic information with clinical outcomes; detects significant genomic and imaging markers associated with clinical prognosis. Our methods are motivated by and applied to the TCGA-based Glioblastoma dataset, through the application of Bayesian continuous shrinkage method, the model identifies important magnetic resonance imaging features and the associated genomic platforms that significantly affect patients' survival.

EMAIL: youyimimi66@gmail.com

45g. Comparing Test-Retest Reliability of Dynamic Functional Connectivity Methods

Martin A. Lindquist, Johns Hopkins University
Yuting Xu*, Johns Hopkins University
Brian Caffo, Johns Hopkins University

Due to the dynamic nature of brain activity, interest in estimating the rapid functional connectivity (FC) changes that occur during resting-state fMRI has recently soared. However, studying dynamic FC is challenging: due to the low signal-to-noise ratio in fMRI, metrics of dynamic FC are sensitive to the estimation method used. Moreover, in contrast to many statistical methods that provide data reduction, standard dynamic FC methods substantially increase the initial number of data points for analysis. Thus, it is critical to establish methods and summary metrics that maximize reliability and the utility of dynamic FC to provide insight into brain function. We investigated the reliability of three commonly used estimation methods for dynamic FC: sliding window, tapered sliding window, and dynamic conditional correlations. The reliability of these methods was compared using publicly available rs-fMRI test-retest data sets. We find that dynamic correlations are reliably detected in those test-retest data sets.

EMAIL: xuyuting1990@gmail.com

45h. Supervised Multiway Factorization

Eric F. Lock*, University of Minnesota
Gen Li, Columbia University

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 describe a 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.

EMAIL: elock@umn.edu

46. Gene Networks

› Error Quantification in Biologically Relevant eQTL Network Metrics

Sheila Gaynor*, Harvard University
Maud Fagny, Dana-Farber Cancer Institute
Megha Padi, Dana-Farber Cancer Institute
John Platig, Dana-Farber Cancer Institute
John Quackenbush, Harvard University and Dana-Farber Cancer Institute

Expression quantitative trait locus (eQTL) analysis associates SNPs with genes by regressing genotype on the continuous measure of gene expression. Graphical representation of significant eQTL SNP-gene associations and the structural analysis of the resulting graph must provide insight into the complex SNP-gene associations influencing biological processes including disease. However, these graphical representations treat eQTL associations as discrete, ignoring the variability that comes from the eQTL regression estimates, and failing to account for the effect of this variability on the network, its structure, and the conclusions that can be drawn from them. We have developed a method to incorporate these errors estimates into the calculation of network metrics such as degree centrality, as well as more complex structural properties such as community structure. We demonstrate our approach using a network representation of eQTL associations estimated from a cohort study of chronic obstructive pulmonary disease (COPD). We show that incorporating error estimates can aid in the identification of biologically-relevant associations and allow better replication of network metrics.

EMAIL: sgaynor@fas.harvard.edu

Bayesian Nonparametric Feature Classification over Large-Scale Gene Networks with Missing Values

Zhuxuan Jin*, Emory University
Zhou Lan, North Carolina State University
Jian Kang, University of Michigan
Tianwei Yu, Emory University

Selecting and classifying features over large-scale gene networks has become increasingly important. However, most recent works on gene selection cannot distinguish specific types of effects, lose selection accuracy and introduce bias in estimating gene effects. To address those limitations, we propose a Bayesian nonparametric method for gene and sub-network selection and classification. It can classify important genes with two different behaviors: "down-regulated" and "up-regulated" for which a new prior model is developed for the class indicator incorporating the network dependence. In posterior computation, we resort to fully Bayesian inference incorporating Swendsen-Wang algorithm for efficiently updating class indicators. The proposed method can handle missing data, which improves the selection and classification accuracy and reduces the bias in estimating gene effects. We illustrate our methods in simulation studies and the analysis of the cutaneous melanoma dataset from the Cancer Genome Atlas.

EMAIL: zjin23@emory.edu

Composite Likelihood Inference on Big Genomic Network

Ningtao Wang*, University of Texas Health Science Center at Houston

Hi-C experiment allows the measurements of the frequencies of physical contacts among pairs of genomic loci at a genome-wide scale. A fundamental step of analyzing such data is to detect and model the community structure, where the stochastic block model is the most commonly used benchmark for such a task. The first problem of fitting stochastic block model is the computation of optimal label assignments of communities, which is, in principle, NP-hard. We proposed to use composite likelihoods as an approximation of the infeasible likelihood. A two-layer EM algorithm has been developed to obtain the parameter estimates and converts the NP-hard problem into a linear-hard one. We proved the consistency of the latent class assignment function when the number of observations tends to infinity. The model was validated through simulation studies and by new discoveries from analyzing Hi-C data.

EMAIL: Ningtao.Wang@uth.tmc.edu

Adaptive Testing of SNP-Brain Functional Connectivity Association via a Modular Network Analysis

Chen Gao*, University of Minnesota
Junghi Kim, University of Minnesota
Wei Pan, University of Minnesota

Due to its high dimensionality and high noise levels, analysis of a large brain functional network may not be powerful and easy to interpret. Although several methods exist for estimating brain functional networks, it is still difficult to extract modules from such network estimates. Motivated by these considerations, we adapt a weighted gene co-expression network analysis framework to resting-state fMRI data to identify modular structures in brain functional networks. Modular structures are identified by using topological overlap matrix elements in hierarchical clustering. We propose applying a new adaptive test built on the proportional odds model that can be applied to a high-dimensional setting, where the number of variables (p) can exceed the sample size (n) in addition to the usual p smaller than n setting. We applied our proposed methods to the ADNI data to test for associations between a genetic variant and either the whole brain functional network or its various subcomponents using various connectivity measures. We uncovered several modules based on the control cohort, and some of them were associated with the APOE4 variant and several other SNPs.

EMAIL: gaoxx492@umn.edu

› Bayesian Variable Selection over Large Scale Networks via the Thresholded Graph Laplacian Gaussian Prior with Application to Genomics

Qingpo Cai*, Emory University
Jian Kang, University of Michigan
Tianwei Yu, Emory University

Selecting informative features from tons of thousands of candidate genes becomes increasingly important in current genomic research. A promising approach is to perform variable selection under regression models while incorporating the existing biological structural information. Most existing methods focus on the local network structure and require heavy computational costs for the large scale problem. In this work, we propose a novel prior model for Bayesian variable selection over large scale networks in the generalized linear model (GLM) framework: the Thresholded Graph Laplacian Gaussian (TGLG) prior, which adopts the graph Laplacian matrix to characterize the conditional dependence between neighboring predictors accounting for the global network structure. Under mild conditions, we show the proposed model enjoys the posterior consistency. We also develop a Metropolis-adjusted Langevin algorithm (MALA) for efficient posterior computation. We illustrate the superiorities of the proposed method compared with existing alternatives via extensive simulation studies and an analysis of the melanoma gene expression dataset in the Cancer Genome Atlas (TCGA).

EMAIL: qingpo.cai@emory.edu

› Normalizing Nonlinear Single-Cell Sequencing Data with ERCC Spike-in Genes

Nicholas Lytal*, University of Arizona
Lingling An, University of Arizona

Single-cell RNA sequencing (scRNA-seq) has lead to significant advances in identifying rare cell types and heterogeneity within cell groups, both of which bulk sequencing methods fail to detect. However, restrictions such as amplification bias and technical noise often limit the power of scRNA-seq results. To address the issue of technical noise, various normalization methods exist that utilize the known expression values of exogenous ERCC spike-in genes to build a prediction model. Although several such methods exist, many involve prediction models that are not well equipped to normalize scRNA_seq data that follows a nonlinear pattern. We propose a novel scRNA-seq normalization method that uses a form of partial least squares to effectively identify and adjust for technical variation even for scRNA-seq data with nonlinear patterns, using data simulation to support our results.

EMAIL: njlytal@email.arizona.edu

47. Generalized Linear Models

› Quasi-Likelihood Ratio Tests for Homoscedasticity in Linear Regression

Lili Yu, Georgia Southern University
Varadan Sevilimedu*, Georgia Southern University
Robert Vogel, Georgia Southern University
Hani Samawi, Georgia Southern University

It is important to check for homoscedasticity in regression models. The violation of this assumption can lead to inefficient estimation or incorrect inference. The tests that have been proposed for homoscedasticity usually require that the errors are normally distributed or that the variance function is parametric. These assumptions greatly restrict their application to real data analysis. In this paper, we propose two quasi-likelihood ratio (QLR) tests with minimum assumptions for linear regression models i.e. they do not require a known distribution of the error terms and they don't require parametric variance functions. In addition, they can be extended to complex non-linear and non-parametric models with the same minimal assumptions. This study provides the details of constructing the two QLR tests, their asymptotic properties, their application via simulation using the MASS package in R and real data analysis relating to these tests.

EMAIL: varadan.sevilimedu@gmail.com

Method of Divide-and-Combine in Regularized Generalized Linear Models for Big Data

Lu Tang*, University of Michigan
Ling Zhou, University of Michigan
Peter X.K. Song, University of Michigan

When a dataset is too big to be analyzed entirely once by a single computer, the strategy of divide-and-combine has been the method of choice to overcome the computational hurdle due to its scalability. Although random data partition has been widely adopted, there is lack of clear theoretical justification and practical guidelines to combine results obtained from separate analysis of individual sub-datasets, especially when a regularization method such as lasso is used for variable selection to improve numerical stability. In this paper we develop a new strategy to combine separate lasso-type estimates of regression parameters by the confidence distributions based on bias-corrected estimators. We first establish the approach to construct the confidence distribution and then show that the resulting combined estimator enjoys the Fisher's efficiency in the sense of the estimation efficiency achieved by the maximum likelihood estimator from full data. Furthermore, using the combined regularized estimator we propose an inference procedure. Extensive simulation studies are presented to evaluate the performance of the proposed methodology with comparisons to competing methods.

EMAIL: lutang@umich.edu

A Method for Inter-Modal Segmentation Analysis

Alessandra M. Valcarcel*, University of Pennsylvania
Russell Shinohara, University of Pennsylvania

Magnetic resonance imaging (MRI) is crucial for in vivo detection and characterization of white matter lesions (WML) in multiple sclerosis. While these lesions have been studied for over two decades using MRI technology, the accurate automated detection and delineation of WMLs remains challenging. The majority of statistical techniques for the automated segmentation of WML are based on a single imaging modality. However, recent advances have centered on multimodal techniques for identifying WML using mean modeling. These complementary modalities emphasize different tissue properties which can help identify and characterize different features of lesions. To harness the coherent changes in these measurements, we propose to utilize inter-modal coupling regression to estimate the covariance structure across modalities. We then use a local regression which leverages these new covariance features as well as the mean structure of each imaging modality in order to model the probability that any voxel is part of a lesion. We also introduce a novel thresholding algorithm to fully automate the estimation of the probability maps to generate fully automated segmentations masks.

EMAIL: alval@mail.med.upenn.edu

Poisson Regression Models for Zero and K Inflated Count Data

Monika Arora*, Old Dominion University
N. Rao Chaganty, Old Dominion University

Several examples of count data consists of inflated count zero. There are numerous papers in the literature that show how to fit Poisson regression models that account for the zero inflation. In this research we study Poisson regression models for data that consists of inflated frequency of some positive k in addition to zero inflation. We derive basic properties of the model and describe how to obtain the maximum likelihood estimate of the regression parameter using EM algorithm. We derive the observed information matrix which yields the standard errors of the EM estimates using the ideas presented in Louis (1982). Two real life examples will be used to illustrate the procedure of fitting our regression models.

EMAIL: maror001@odu.edu

A Unified Framework for Testing the Fixed and Random Effects Jointly in the Generalized Linear Mixed Models

Jingchunzi Shi*, University of Michigan
Seunggeun Lee, University of Michigan

The framework of testing the fixed and random effects jointly is of considerable applications in biomedical studies. One application example is to use such tests for ascertaining associations when there exists heterogeneity in GWAS

meta-analysis; another example is the nonparametric test of spline curves. Although extensive research was done on testing random effect terms only, little work was developed for the joint test of fixed and random effects. Here, we propose a score test for the joint test in GLMMs with one variance component. Our method first reparameterizes fixed effects terms as a product of a scale parameter and a vector of nuisance parameters. With such reparameterization, joint testing for the fixed and random effects is equivalent for testing whether the scale parameter is zero. Since the nuisance parameters are hidden under the null hypothesis, we propose to use the supremum of score tests over the nuisance parameters as our test statistic. P-values can be obtained either analytically or using the monte-calro algorithm. We investigated performances of our method through simulation studies and real data application.

EMAIL: shijingc@umich.edu

› Measuring the Individual Benefit of a Medical or Behavioral Treatment using Generalized Linear Mixed-Effects Models

Francisco J. Diaz*, University of Kansas Medical Center

We present statistical definitions of the individual benefit of a medical or behavioral treatment and of the severity of a chronic illness. These definitions are used to develop a graphical method that can be used by statisticians and clinicians in the data analysis of clinical trials from the perspective of personalized medicine. The method focuses on assessing and comparing individual effects of treatments rather than average effects, and can be used with continuous and discrete responses, including dichotomous and count responses. The method is based on recently published developments in generalized linear mixed-effects models. To illustrate, analyses of data from the STAR*D clinical trial of sequences of treatments for depression and data from a clinical trial of respiratory treatments are presented. The estimation of individual benefits is also explained.

EMAIL: fdiaz@kumc.edu

48. Personalized Medicine

› Distributed Learning from Multiple EHR Databases: Contextual Embedding Models for Medical Events

Ziyi Li*, Emory University
Qi Long, Emory University
Xiaoqian Jiang, University of California, San Diego

Electronic health record(EHR) data provide promising opportunities to explore personalized treatment regime. Compared with other clinical data, EHR data are known for their irregularity and complexity. In addition, analyzing EHR data involves privacy issues and sharing such data is often infeasible among research sites. A recent work by Farhan et al. (2016) successfully builds one predictive model for more than seventy common diagnoses. Although it achieves a relatively high predictive accuracy, the current model cannot build global models without sharing data among sites. In this work, we proposed three novel contextual embedding methods for diagnoses prediction: Naive updates, Dropout updates, and Distributed Noise Contrastive Estimation (NCE). We also extend Distributed NCE with Differential Privacy to obtain reliable data privacy protections. Our simulation study with a real dataset demonstrates that the proposed methods not only can build predictive model with privacy protection, but also well preserve the model structure and achieve comparable prediction accuracy compared with gold standard model built with all the data.

EMAIL: ziyi.li@emory.edu

› Single-Index Models for Personalized Medicine

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 flexible single-index model to allow complex treatment-covariate interactions and to derive a simple linear tratement rule for personalized treatment decision. Our model extends the proportional hazards models by using a monotone regression function. The inference procedures are based on sieve estimation that includes single-index parameters in the basis expansion. We obtain the theoretical results by deriving

the necessary rates of convergence for the nonparamatric estimator of the arbitrary regression function. We provide simuntaneous inference on single-index parameters and other regression parameters. Simulation studies are conducted to assess the finite sample performance. An application to a multiple-type cancer study is presented to illustrate our methods.

EMAIL: jinjin@live.unc.edu

› New Covariate-Adjusted Response-Adaptive Designs for Personalized Medicine

Wanying Zhao*, The George Washington University
Feifang Hu, The George Washington University

To develop personalized medicine, more covariates are under consideration in a clinical trial. According to the different roles played in clinical studies, covariates can be classified as prognostic covariates and predictive covariates. In literature, no design is available to deal with them simultaneously. We establish an innovative class of covariate-adjusted response-adaptive (CARA) designs, which incorporate both prognostic and predictive covariates in the randomization procedure, and provide a theoretical foundation of hypothesis testing under the new CARA design based on linear models. We derive the asymptotic distributions of the test statistics of testing both treatment effects and the significance of covariates under null and alternative hypotheses. Under the new CARA design, the hypothesis testing to compare treatment effects is usually conservative, but more powerful than complete randomization. Numeric studies with different scenarios are provided to illustrate the theoretical conclusions.

EMAIL: zhaowanying90@gmail.com

› Enhanced Statistical Models and Internet-Based Communication Tools for Precision Cancer Risk Prediction

Donna Pauler Ankerst*, Technical University Munich and University of Texas Health Science Center at San Antonio
Martin Goros, University of Texas Health Science Center at San Antonio

As advances in the understanding of cancer have emerged, so too have treatment options diverged, necessitating more intricate discussions with the patient. For example, cancer is now rarely envisioned as a zero or one outcome, but rather in terms of level of aggressiveness. This mandates a move towards multivariate outcome risk prediction accompanied by novel patient-accessible visualization. The easy-to-use R Shiny package allows statisticians to seamlessly demonstrate complex models to clinical collaborators, thereby facilitating valuable subject matter input. Posting the R Shiny interface online bridges the gap between academic research and the patients it was designed to help, in addition to facilitating external validation. We and others have continually stressed the importance of dynamic models that are constantly updated as new data become available, new markers are discovered, or changes in diagnostic technology are witnessed. We illustrate these concepts with our own experience building and maintaining the Prostate Cancer Prevention Trial Risk Calculator Version 2.0, available at myprostatecancerrisk.com.

EMAIL: ankerst@tum.de

› Optimal Treatment Estimation in Additive Hazards Model

Suhyun Kang*, North Carolina State University
Wenbin Lu, North Carolina State University

We propose a doubly robust estimation method for the optimal treatment regime based on an additive hazards model with censored survival data. Specifically, we introduce a new semiparametric additive hazard model which allows flexible baseline covariate effects in the control group and incorporates marginal treatment effect and its linear interaction with covariates. In addition, we propose a time-dependent propensity score to construct an A-learning type of estimating equations. The resulting estimator is shown to be consistent and asymptotically normal when either the baseline effect model for covariates or the propensity score is correctly specified. The asymptotic variance of the estimator is consistently estimated using a simple resampling method. Simulation studies are conducted to evaluate the finite-sample performance of the estimators and an application to AIDS clinical trial data is also given to illustrate the methodology.

EMAIL: skang8@ncsu.edu

49. Prediction/Prognostic Modeling

Predicting Human Driving Behavior to Help Driverless Vehicles Drive: Random Intercept BART

Yaoyuan Vincent Tan*, University of Michigan
Carol A.C. Flannagan, University of Michigan
Michael R. Elliott, University of Michigan

Prediction models range from simple models like the linear regression to more sophisticated models like the multivariate adaptive regression splines and classification regression trees. Bayesian additive regression trees (BART) works well in a variety of situations because it is able to incorporate non-linear and complex interaction effects easily. Unfortunately, BART was developed under the assumption of independence between subjects and hence, cannot be applied to studies that have longitudinal or clustered outcomes. Although there have been attempts to extend BART to correlated outcomes, the models proposed were either too complicated or cannot be easily implemented. In this paper, we extend BART to correlated outcomes by adding a random intercept (riBART). Compared to previous extensions, our method is simple to implement. We consider the repeated sampling properties of riBART in a simulation study, and apply our method to predict whether a human driven vehicle would stop before executing a left turn at intersections. Such a model would aid driverless vehicles in reducing the number of accidents with human driven vehicles at intersections.

EMAIL: vincetan@umich.edu

A Default Prior for the Intercept in Binary-Data Regression Models

Philip S. Boonstra*, University of Michigan
Ananda Sen, University of Michigan

A dataset with a binary outcome (Y) is "completely separated" when a linear combination of the covariates (X) perfectly discriminates those observations falling in one category of the outcome from those falling in the other. Separation may occur when the dimension of X is close the length of Y, increasing variability in both parameter estimates and predictions. In the Bayesian framework, default shrinkage priors have been proposed for regression parameters to reduce this variability. However, relatively little focus has been placed on the choice of prior for the intercept parameter. Typically, a diffuse or improper flat prior is placed directly on the intercept. Motivated by a simple algebraic result demonstrating the importance of estimating the intercept well in cases of (near) separation, we suggest an alternative default prior that is flat on the probability scale when the covariates are at their expectation: $\Pr(Y=1 \mid X=EX) \sim \text{Unif}(0,1)$. We demonstrate, by way of a simulation study and analysis of a biomedical dataset for classification with many predictors, the possible improvement in both estimation and prediction accuracy compared to typical choices of prior.

EMAIL: philb@umich.edu

Development of an Improved Comorbidity Measure Based on the ICD System Using Statistical and Machine Learning Methods

Ralph Ward*, Medical University of South Carolina
Leonard Egede, Medical University of South Carolina
Viswanathan Ramakrishnan, Medical University of South Carolina
Lewis Frey, Medical University of South Carolina
Robert Neal Axon, Medical University of South Carolina
Mulugeta Gebregziabher, Medical University of South Carolina

In health outcomes research, adjustment for disease burden is essential in order to help minimize the potential for bias. Disease burden is often measured by summarizing binary diagnostic codes data such as those used in the International Classification of Diseases system. Many researchers rely on well-established ICD summary measures like the Elixhauser Index, but often these do not account for disease severity or possible interactions between conditions. In this paper, we compared several methods for modeling binary ICD data and showed that several produced more accurate predictions than a model based on the Elixhauser index. We used area under the receiver operator characteristics curve (AUC), net reclassification improvement (NRI) statistics and the Brier score to compare models. Overall, the Bayesian additive regression trees (BART), random forest, elastic-net, and pooled models had superior predictive performance. We tested these mod-

els on a pooled dataset from three large VA chronic disease cohorts (diabetes, chronic kidney disease and traumatic brain injury) characterized by a wide variety of comorbidities.

EMAIL: rcccward@gmail.com

› Prediction Models for Network-Linked Data

Tianxi Li*, University of Michigan
Elizaveta Levina, University of Michigan
Ji Zhu, University of Michigan

Prediction problems typically assume the training data are independent samples, but in many modern applications samples come from individuals connected by a network. For example, in adolescent health studies of risk-taking behaviors, information on the subjects' social networks is often available and plays an important role through network cohesion, the empirically observed phenomenon of friends behaving similarly. Taking cohesion into account in prediction models should allow us to improve their performance. Here we propose a regression model with a network-based penalty on individual node effects to encourage similarity between predictions for linked nodes, and show that it performs better than traditional models both theoretically and empirically when network cohesion is present. The framework is easily extended to other models, such as the generalized linear model and Cox's proportional hazard model. Applications to predicting levels of recreational activity and marijuana usage among teenagers based on both demographic covariates and their friendship networks are discussed in detail and demonstrate the effectiveness of our approach.

EMAIL: tianxili@umich.edu

› Program Chairs vs. Machine: A New Approach to ENAR Program Creation

Jiaqi Li*, Mathematica Policy Research
Fei Xing, Mathematica Policy Research
Mariel Finucane, Mathematica Policy Research
Scarlett Bellamy, Drexel University
Andrea Foulkes, Mount Holyoke College
Nandita Mitra, University of Pennsylvania

Have you ever wondered how the ENAR scientific program materializes? Have you ever felt that your talk has been placed in a session where it does not belong? Traditionally, the ENAR Program Chair and Associate Program Chair receive hundreds of contributed abstracts which they collate "by hand" into about 45 sessions of similar topic. This year we decided to explore a new and more modern data science approach. We applied text mining to extract key information from the title, abstract, and key words of each submission, and then integrated modern machine learning and Bayesian techniques to categorize each talk into an appropriate session. We first determined the consistency and overlap of the Program Chairs' designated sessions to those formed by the computational approach. We then sent out surveys to a random selection of authors for whom the 'human" and "machine" approaches made different designations, asking them to choose which of the session placements was more appropriate for their talk. We hope this exercise provides some guidance for using machine learning tools in this and similar practical settings.

EMAIL: JLi@mathematica-mpr.com

› Informing a Risk Prediction Model for Binary Outcomes with External Coefficient Information

Wenting Cheng*, University of Michigan
Jeremy M. G. Taylor, University of Michigan
Bhramar Mukherjee, University of Michigan

We consider a situation where there is rich historical data available for the coefficients and their standard errors in an established regression model $Pr(Y = 1|X)$, from a large study. We would like to utilize this summary information for improving estimation and prediction in an expanded model of interest, $Pr(Y=1|X, B)$. The additional variable B is a new biomarker, measured on a small number of subjects in a new dataset. We develop and evaluate several approaches for translating the external information into constraints on regression coefficients. Borrowing from the measurement error literature we establish an approximate relationship between the regression coefficients in the models $Pr(Y = 1|X, \beta)$, $Pr(Y = 1|X, B, \gamma)$ and $E(B|X, \theta)$ for a Gaussian distribution of B. For binary B we propose an alternate expression. The simulation results comparing these methods indicate that

historical information on $Pr(Y = 1|X, \beta)$ can improve the efficiency of estimation and enhance the predictive power in $Pr(Y = 1|X, B, \gamma)$. We illustrate our methodology by enhancing the High-grade Prostate Cancer Prevention Trial Risk Calculator, with a new biomarker PCA3.

EMAIL: chengwt@umich.edu

50. Methods for Next Generation Sequencing Data

› A Comparison Study of Multivariate Fixed Models and Gene Association with Multiple Traits (GAMuT) for Next-Generation Sequencing

Ruzong Fan*, Georgetown University
Jeesun Jung, National Institute on Alcohol Abuse and Alcoholism, National Institutes of Health
Chi-yang Chiu, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health
Daniel E. Weeks, University of Pittsburgh
Alexander F. Wilson, National Human Genome Research Institute, National Institutes of Health
Joan E. Bailey-Wilson, National Human Genome Research Institute, National Institutes of Health
Christopher I. Amos, Dartmouth Medical College
Michael Boehnke, University of Michigan
Momiao Xiong, University of Texas Health Science Center at Houston

In this project, extensive simulations are performed to compare two statistical methods to analyze multiple correlated quantitative phenotypes: (1) approximate F-distributed tests of multivariate functional linear models (MFLM) and additive models of multivariate analysis of variance (MANOVA), and (2) Gene Association with Multiple Traits (GAMuT) for association testing of high-dimensional genotype data. It is shown that approximate F-distributed tests of MFLM and MANOVA have higher power and are more appropriate for major gene association analysis (i.e., scenarios in which some genetic variants have relatively large effects on the phenotypes); GAMuT has higher power and is more appropriate for analyzing polygenic effects (i.e., effects from a large number of genetic variants each of which contributes a small amount to the phenotypes). MFLM and MANOVA are very flexible and can be used to perform association analysis for: (i) rare variants, (ii) common variants, and (iii) a combination of rare and common variants.

EMAIL: rf740@georgetown.edu

› Incorporating Linkage Disequilibrium and Identity-by-Descent Information to Improve Genotype Calling from Family-Based Sequencing Data

Zhou Fang*, University of Pittsburgh
Qi Yan, Children's Hospital of Pittsburgh of UPMC
George C. Tseng, University of Pittsburgh
Bingshan Li, Vanderbilt University Medical Center
Wei Chen, University of Pittsburgh and Children's Hospital of Pittsburgh of UPMC

Next generation sequencing technologies empowered the genome-wide discovery of genetic variants in great detail. Previous studies showed that when pedigree information is wisely leveraged, sequencing more individuals at shallower coverage is more cost-effective. Current genotype calling methods retrieve genotyping accuracy using either linkage disequilibrium (LD) or identity-by-descent (IBD) information. While neither of the two types of information shows universal advantages over the other across variants of different rarity, no method exists to accommodate the two types of information simultaneously. Here, we propose a novel genotype inference method to integrate both LD and IBD information. To evaluate our approach, we simulated individuals in different family structures, with variants of all rarity sequenced in a wide range of depth. Simulation results showed that with an informative family structure, our information-duo method could significantly increase the genotype accuracy. Furthermore, it displays consistent lead in accuracy over single-information methods for both rare and common variants. We implemented this method in a user-friendly package using C++.

EMAIL: fangz.ark@gmail.com

Using Bayes Factor to Detect Differentially Methylated Regions Associated with Disease Severity

Fengjiao Hu*, National Institute of Environmental Health Sciences, National Institutes of Health
Hongyan Xu, Augusta University
Duchwan Ryu, Northern Illinois University
Santu Ghosh, Augusta University
Huidong Shi, Augusta University
Varghese George, Augusta University

Researchers in genomics are increasingly interested in epigenetic factors such as DNA methylation, because they play an important role in regulating gene expression without changes in the DNA sequence. There have been significant advances in recent years in developing statistical methods to detect DMRs associated with binary disease status. In our approach, we consider multiple severity levels of disease, and develop a Bayesian statistical method to detect the region with increasing (or decreasing) methylation rates as the disease severity increases (or decreases). Patients are classified into more than two groups, based on the disease severity, and DMRs are detected by using the moving windows along the genome. Within each window, Bayes factor is calculated to test the hypothesis of monotonic increase in methylation rates corresponding to disease severity versus no difference. A mixed-effect model is used to incorporate the correlation of methylation rates of nearby CpG sites. Results from extensive simulation indicate that our proposed method is statistically valid and reasonably powerful. We demonstrate our approach on a bisulfite sequencing dataset from a CLL study.

EMAIL: fengjiao.hu@nih.gov

PhredEM: A Phred-Score-Informed Genotype-Calling Approach for Next-Generation Sequencing Studies

Peizhou Liao*, Emory University
Glen A. Satten, Centers for Disease Control and Prevention
Yijuan Hu, Emory University

A fundamental challenge in analyzing NGS data is to determine an individual's genotype accurately. Correctly estimating base-calling error rates is critical to accurate genotype calls. Phred scores can be used to decide which calls are reliable. Some genotype callers like GATK, directly calculate the base-calling error rates from phred scores or recalibrated base quality scores. Others like SeqEM, estimate error rates from the read data only. We propose a new likelihood-based genotype-calling approach that exploits all reads and estimates the per-base error rates by incorporating phred scores through a logistic regression model. Our approach (PhredEM), uses the EM algorithm to obtain consistent parameter estimates. It also includes a simple, computationally efficient screening algorithm. PhredEM can be used together with a LD-based method to further improve genotype calling. We evaluate the performance of PhredEM using simulated data and real sequencing data from the UK10K project and 1000 Genomes project. The results demonstrate that PhredEM is an improved, robust and widely applicable genotype-calling approach for NGS studies.

EMAIL: pliao3@emory.edu

A Statistical Method for the Analysis of Multiple ChIP-Seq Datasets

Pedro Luiz Baldoni*, University of North Carolina, Chapel Hill
Naim Rashid, University of North Carolina, Chapel Hill
Joseph G. Ibrahim, University of North Carolina, Chapel Hill

Chromatin immunoprecipitation followed by next generation sequencing (ChIP-seq) is an experimental technique to detect regions of protein-DNA interaction in the genome, such a transcription factor binding sites or regions containing histone modifications ("enrichment regions"). A common 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 over time, novel methods to detect consensus regions of enrichment across multiple samples or differential enrichment between groups of samples are of interest. We develop a statistical method and software to simultaneously analyze multiple histone modification ChIP-seq datasets through a class of efficient Mixed Hidden Markov Models (MHMM) to call consensus broad regions region of enrichment found across samples. This novel methodology will provide a tool able to compare data from different experiments in a single framework by considering subject and

population-specific variation in a single model. Simulation studies and real data results will be presented.

EMAIL: baldoni@email.unc.edu

51. Combining Data From Multiple Sources: Applications From Health Policy

› Addressing Differential Measurement Error in Self-Reported Dietary Data using an External Validation Study: Application to a Longitudinal Lifestyle Intervention Trial

Juned Siddique*, Northwestern University Feinberg School of Medicine
Laurence Freedman, Gertner Institute for Epidemiology and Health Policy Research
Raymond Carroll, Texas A&M University
Trivellore Raghunathan, University of Michigan
Elizabeth Stuart, Johns Hopkins Bloomberg School of Public Health

In lifestyle intervention trials, where the goal is to change a participant's weight or modify their eating behavior, self-reported diet is a longitudinal outcome variable that is subject to measurement error. We propose a statistical framework for correcting for measurement error in longitudinal self-reported dietary data by combining these data with auxiliary data from external biomarker validation studies where both self-reported and recovery biomarkers of dietary intake are available. In this setting, dietary intake measured without error in the intervention trial is treated as missing data and multiple imputation is used to fill-in the missing measurements. Since validation studies are cross-sectional, they do not contain information on whether the nature of the measurement error changes over time or differs between treatment and control groups. We use sensitivity analyses to address the influence of these unverifiable assumptions involving the measurement error process and how they affect inferences regarding the effect of treatment.

EMAIL: siddique@northwestern.edu

› Multilevel Imputation for Missing Values in Meta-Analysis with Individual Participant Data

Yajuan Si*, University of Wisconsin, Madison

Meta-analysis with individual participant data (IPD-MA) has become increasingly popular for comparative effectiveness research. However, missing data problems in IPD-MA become challenging. Different measures could be used to assess the same construct in different studies. Some covariates were omitted across studies. Nonetheless, the hierarchical structure of IPD-MA is often ignored when handling missing data. We develop new statistical methods to handle the missing values in IPD-MA and evaluate the effect on the subsequent analysis. We have identified patients of locoregional invasive breast cancer participating in 17 legacy Alliance trials. Our scientific goal is to examine the likelihood and timing of recurrence and delayed treatment toxicity across the spectrum of risk experienced by breast cancer patients treated with curative intent. We extend multiple imputation approaches to handle the missing outcome measures and covariates. The proposed hierarchical approach will capture the heterogeneity and yield robust estimates as demonstrated by simulation studies and the application.

EMAIL: ysi@biostat.wisc.edu

› Combining Item Response Theory with Multiple Imputation to Equate Health Assessment Questionnaires

Chenyang Gu, Harvard University
Roee Gutman*, Brown University
Vincent Mor, Brown University

Assessment of patients' functional status across the continuum of care requires a common instrument. Different health care settings rely on different instruments. The Functional Independence Measure (FIM) is used to evaluate the functional status of patients who stay in inpatient rehabilitation facilities (IRFs). After discharge from IRFs, the Minimum Data Set (MDS) is collected for patients that are transferred to skilled nursing facilities (SNFs), while the Outcome and Assessment Information Set (OASIS) is collected for patients receiving home health care. To compare patients that are discharged from IRFs to either SNFs or home health, a single

measure of functionality is required. Assuming that all the patients have observed FIM measurements and treating the unmeasured MDS or OASIS items as missing, we propose a variant of the predictive mean matching method, which relies on Item Response Theory to impute the unobserved functionality items. Using simulations, we compared the proposed approach to existing methods, and showed that it performs well for estimating the overall functionality status, while preserving the correlation structures among functionality items.

EMAIL: roee_gutman@brown.edu

52. Platform Trial Designs: Benefits, Efficiencies, and Applications

› Efficiencies of Multi-Arm Platform Clinical Trials

Benjamin R. Saville*, Berry Consultants
Scott Berry, Berry Consultants

A "platform trial" is a clinical trial with a single master protocol in which multiple treatments are evaluated simultaneously. Adaptive platform designs offer flexible features such as dropping treatments for futility, declaring one or more treatments superior, or adding new treatments to be tested during the course of a trial. Such designs can be more efficient at finding beneficial treatments relative to traditional two group designs. We quantify these efficiencies via simulation to show that platform trials can find beneficial treatments with fewer patients, fewer patient failures, less time, and with greater probabilities of success than traditional two-arm strategies. In an era of personalized medicine, platform trials provide the innovation needed to efficiently evaluate modern treatments.

EMAIL: ben@berryconsultants.com

› Platform Trials Designed to Find Disease Modifying Therapies for Alzheimer's Disease

Melanie Quintana*, Berry Consultants
Scott Berry, Berry Consultants

We describe the development of a platform trial for determining the effectiveness of therapies for Alzheimer's Disease (AD). AD is the leading cause of dementia globally. Current approved therapies have only been shown to slow the worsening of symptoms but not alter the disease progression. Our primary interest is in developing a platform trial with the infrastructure that enables simultaneous evaluation of multiple potential disease modifying therapies in individuals at risk for AD. The trial follows a two-stage design. The goal of the first stage is the demonstration of target biomarker modulation. The goal of the second stage is the demonstration of efficacy in terms of a cognitive benefit. Several innovative aspects of the design include: pooling of information from placebo patients across treatment regimens, proposing to stop a regimen early for lack of biomarker or cognitive efficacy, and utilizing disease progression models to determine if a treatment is effective in slowing the rate of cognitive decline. Within our presentation we highlight the importance of using natural history data to guide disease modeling and clinical trial simulation.

EMAIL: melanie@berryconsultants.com

› Platform Trials for Patients with Glioblastoma

Brian M. Alexander*, Dana-Farber Cancer Institute and Harvard Medical School

Platform trials have the capacity to reorganize therapeutic development around the patient and specific diseases. Glioblastoma is a highly deadly brain cancer with no early detection, no prevention, and minimal effective therapies. This session will discuss trial design and implementation of biomarker-based, Bayesian response adaptively randomized platform trials for glioblastoma, including the GBM Adaptive Global Innovative Learning Environment (AGILE). Specific trial design elements and decisions will be discussed as will the other opportunities that creating such structures generates with respect to patient engagement, funding, and regulatory science.

EMAIL: brian_alexander@dfci.harvard.edu

53. Statistical Causal Inference Using Propensity Scores for Complex Survey Data Sets

› Statistical Causal Inference of Estimating the Effects of Latent Classes

Joseph Kang*, Centers for Disease Control and Prevention
Jaeyoung Hong, Centers for Disease Control and Prevention
Yulei He, Centers for Disease Control and Prevention
Tianqi Chen, Dana-Farber Cancer Institute

In the literature of behavioral sciences, latent class analysis (LCA) has been used to effectively cluster multiple survey items. Statistical inference with an exposure variable, which is identified by the LCA model, is challenging because 1) the exposure variable is essentially missing and harbors the uncertainty of estimating parameters in the LCA model and 2) confounding bias adjustments need relevant propensity score models. In addition to these challenges, complex survey design features and survey weights will have to be accounted for if they are present. Our solutions to these issues are to 1) assess point estimates with the design-based estimating function approach which was described in Binder (1983) and 2) obtain variance estimates with the Jackknife technique. Using the CDC's NHANES data set, our LCA model identified a latent class for men who have sex with men (MSM) and built new propensity score weights that adjusted the prevalence rates of Herpes Simplex Virus Type 2 (HSV-2) for MSM.

EMAIL: yma9@cdc.gov

› Propensity Score Analysis with Survey Weighted Data

Greg Ridgeway*, University of Pennsylvania
Stephanie A. Kovalchik, Victoria University
Beth Ann Griffin, RAND Corporation
Mohammed U. Kabeto, University of Michigan

Propensity score analysis is a common method for estimating treatment effects, but researchers dealing with data from survey designs are generally not properly accounting for the sampling weights in their analyses. Moreover, recommendations given in the few existing methodological articles on this subject are susceptible to bias. In this presentation we show through derivation, simulation, and a real data example that using sampling weights in the propensity score estimation stage and the outcome model stage results in an estimator that is robust to a variety of conditions that lead to bias for estimators currently recommended in the statistical literature.

EMAIL: gridge@sas.upenn.edu

› Propensity Score Analysis of a Continuous Exposure with a Mixture Model Approach

Jaeyoung Hong*, Centers for Disease Control and Prevention
Tianqi Chen, Dana-Farber Cancer Institute
Joseph Kang, Centers for Disease Control and Prevention

Propensity score models for continuous exposures have been developed in non-survey data sets with unimodal distributions including a single Gaussian normal distribution model. However, when an exposure appears to have a multimodal distribution in complex survey data sets, clustering techniques can be used to group subjects. This presentation will discuss the application of clustering techniques for a continuous exposure in building a novel propensity score model. In particular, the propensity model will be used to analyze the effects of the number of partners of the men who have sex with men (MSM) on sexually transmitted diseases (STD) in NHANES 2001-2014.

EMAIL: lyy1@cdc.gov

› Multiple Imputation Diagnostics using Propensity Score

Yulei He*, Centers for Disease Control and Prevention
Guangyu Zhang, Centers for Disease Control and Prevention
Bo Lu, The Ohio State University
Nat Schenker, Centers for Disease Control and Prevention

Multiple Imputation is a model-based approach to missing data problems. Despite of its popularity in practice, there is a lack of diagnostic tools for the adequacy of imputed data. We consider this problem in a basic setup, in which a single outcome variable is incomplete and all covariates are fully observed. We first estimate response (or nonresponse) propensity scores using the covariate information. We then assess (a) whether the distribution of observed and imputed values are similar

conditional on estimated propensity scores; (b) whether the distribution of covariates corresponding to the observed and missing cases are similar conditional on the observed/imputed outcome variable and estimated propensity scores. Large dissimilarity in either case might suggest certain inadequacy of imputations. In addition, we provide a simple rule that connects diagnostic results with imputation inferences to further assess the usefulness of the imputed data. Simulation studies and a real-data example are used for illustration.

EMAIL: wdq7@cdc.gov

54. CENS INVITED SESSION: Shaping the Future of the Field: Biostatisticians in a Data-Driven World

› Panel Discussion

Jeffrey Leek, Johns Hopkins Bloomberg School of Public Health
Francesca Dominici, Harvard School of Public Health
Ruth Pfeiffer, National Cancer Institute, National Institutes of Health
Christopher J. Miller, AstraZeneca Pharmaceuticals
Jason T. Connor, Berry Consultants

55. Longitudinal Imaging Analysis: Statistical Challenges and Opportunities

› Assessing Treatment Effects on Longitudinal Imaging

Thomas Bengtsson*, Genentech

Assessing potential population level treatment (or progression) effects on longitudinal clinical images is a challenging statistical problem due to high-dimensionality as well as substantial inter- and intra-patient variability. In the context of longitudinal Positron Emission Imaging (PET) imaging of Alzheimer's disease (AD) patients, this work presents a state-space simulation framework to estimate and predict the spatio-temporal spread and progression of protein depositions (eg Amyloid beta or tau) in the brain of AD patients. The primary use of the proposed simulation model is to power randomized clinical trials where PET is used as pharmacodynamic readout of a possible drug effect.

EMAIL: bengtsson.thomas@gene.com

› Relating Multi-Sequence Longitudinal Intensity Profiles and Clinical Covariates in Incident Multiple Sclerosis Lesions

Ciprian Crainiceanu*, Johns Hopkins University

The formation of multiple sclerosis (MS) lesions is a complex process involving inflammation, tissue damage, and tissue repair - all of which are visible on structural magnetic resonance imaging (MRI) and potentially modifiable by pharmacological therapy. We introduce two statistical models for relating voxel-level, longitudinal, multi-sequence structural MRI intensities within MS lesions to clinical information and therapeutic interventions: (1) a principal component analysis (PCA) and (2) function-on-scalar regression models. We characterize the post-lesion incidence repair process on longitudinal, multi-sequence structural MRI from 34 MS patients as voxel-level intensity profiles. For the PCA regression model, we perform PCA on the intensity profiles to develop a voxel-level biomarker for identifying slow and persistent, long-term intensity changes within lesion tissue voxels. The proposed biomarker's ability to identify such effects is validated by two experienced clinicians. In the function-on-scalar regression, both age and distance to the boundary were found to have a statistically significant association with the lesion intensities.

EMAIL: ccrainic@jhsph.edu

› Tensor Generalized Estimating Equations for Longitudinal Imaging Analysis

Xiang Zhang, North Carolina State University
Lexin Li*, University of California, Berkeley
Hua Zhou, University of California, Los Angeles
Dinggang Shen; University of North Carolina, Chapel Hill

Longitudinal neuroimaging studies are rapidly emerging, where brain images are collected on multiple subjects and each at multiple time points. Analysis of such data is particularly important for understanding disease progression and pathologies, but is also challenging. Brain image is in the form of a multidimensional array, or tensor, which possesses both ultrahigh dimensionality and complex structure. Longitudinally repeated images add another layer of complexity. Despite some recent efforts, there exist very few solutions.

We propose tensor generalized estimating equations (GEE) for longitudinal imaging analysis. The GEE approach accounts for intra-subject correlation, whereas an imposed low rank structure on the coefficient tensor effectively reduces the dimensionality. We propose a scalable estimation algorithm, establish the asymptotic properties, investigate sparsity regularization for the purpose of region selection, and demonstrate the proposed method on both simulated and a real data set from the Alzheimer's Disease Neuroimaging Initiative.

EMAIL: lexinli@berkeley.edu

› Methods to Link Repeated Measures of Functional Brain Networks with Clinical CovariateS

Manjari Narayan*, Stanford University
Genevera I. Allen, Rice University and Jan and Dan Duncan Neurological Institute
Adi-Maron Katz, Stanford University
Colleen-Mills Finnerty, Stanford University
Amit Etkin, Stanford University

Gaussian graphical models (GGMs) are popularly used to model brain networks that capture functional interactions between different brain regions or other units of brain function in human neuroimaging. Motivated by an imaging study of Post Traumatic Stress Disorder (PTSD), we seek to uncover how such functional brain networks reflect underlying cognitive deficits. In particular, this study includes both resting fMRI data as well as context dependent fMRI data where subjects perform tasks involving emotional regulation. Thus, we seek to detect whether multiple context dependent changes to network topologies can be linked to symptom severity or diagnosis. Estimation of a single context dependent GGM is challenging given limited scan time. Moreover, imperfect estimates of these networks can compromise subsequent population inference. To address this problem, we extend a procedure, R3, based on resampling, random penalization and random effect statistics to the case of multivariate network statistics. We demonstrate empirical results of our method using simulation studies and on the PTSD dataset.

EMAIL: manjari@alumni.rice.edu

56. Recent Developments in Zero-Inflated Models for Count Data

› Marginalized Bivariate Zero-Inflated Poisson Regression

Habtamu K. Benecha*, National Agricultural Statistics Service, U.S. Department of Agriculture
John S. Preisser, University of North Carolina, Chapel Hill
Kalyan Das, University of Calcutta
Brian Neelon, Medical University of South Carolina

Bivariate counts with excess zeros are sometimes encountered in health research and many other areas. Such counts are often modeled by extending the bivariate Poisson distribution to allow for extra zeros in the two outcomes. While currently available approaches to model bivariate zero-inflated counts commonly estimate regression coefficients for latent parameters specific to unobserved subpopulations, investigators are often interested in making direct inferences about the marginal means of the bivariate count outcome in the overall population. In this paper, we propose a marginalized bivariate zero-inflated Poisson (MBZIP) model that specifies regression coefficients directly to the marginal means of the two correlated count outcomes, allowing straightforward interpretations for the overall effects of exposure variables on the marginal means. Finite sample properties of the marginalized model maximum likelihood estimates are examined in simulation studies. The MBZIP model is applied to dental caries indices for primary and permanent teeth among children from a school-based fluoride mouthrinse study.

EMAIL: habtamu.benecha@nass.usda.gov

› Zero-Inflation in Cytogenetic Radiation Dosimetry

Pedro Puig*, Universitat Autonoma de Barcelona
Amanda Fernandez-Fontelo, Universitat Autonoma de Barcelona
Manuel Higueras, Newcastle University
Elizabeth A. Ainsbury, Public Health England Centre for Radiation, Chemical and Environmental Hazards

Biological dosimetry is essential for prompt determination of the radiation dose received by an exposed individual.

The dose is usually estimated quantifying the damage produced by radiation at cellular level, for instance by counting the number of chromosome aberrations like dicentrics or micronuclei. The distribution of chromosomal aberrations in partial body irradiation scenarios is zero-inflated, because non-irradiated cells contribute an extra amount of zeros in comparison with the distribution of chromosomal aberrations for whole body irradiation. In this talk we introduce the statistical methodology for dose estimation with zero-inflated counts described in the manual of the International Atomic Energy Agency (IAEA, 2011), and we also summarize the recent research led by our team.

EMAIL: ppuig@mat.uab.cat

› A Robust Score Test of Homogeneity for Zero-Inflated Count Data

Nadeesha Mawella, Kansas State University
Wei-Wen Hsu*, Kansas State University
David Todem, Michigan State University
KyungMann Kim, University of Wisconsin, Madison

Score statistics are often used for the test of homogeneity in zero-inflated models for count data. The most cited justification is that it only requires the model fit under the null hypothesis. However, the true null models are often unknown in practice and these statistics can be invalid when the null models are not correctly specified. As an empirical evidence, an intensive simulation study indicates that the sizes of these score tests for zero-inflated Poisson or Negative Binomial models may behave extremely liberal and unstable when the mean function of baseline distribution or the baseline distribution itself is misspecified. In this paper, we propose a new score test of homogeneity for zero-inflated models which is robust to these misspecifications. Technically, the test is developed under the framework of Poisson-Gamma mixture model which can provide a more general framework to incorporate various baseline distributions without specifying the mean function. The empirical performances of our test in finite samples are evaluated in simulation studies and the dental caries data from the Detroit Dental Health Project (DDHP) are used to illustrate the proposed test.

EMAIL: wwhsu@k-state.edu

57. ORAL POSTERS: New Ideas in Causal Inference

57a. INVITED ORAL POSTER:

Estimating the Malaria Attributable Fever Fraction Accounting for Parasites being Killed by Fever and Measurement Error

Kwonsang Lee, University Pennsylvania
Dylan S. Small*, University Pennsylvania

The fraction of fevers that are attributable to malaria, the malaria attributable fever fraction (MAFF), is an important public health measure. Estimating the MAFF is not straightforward because there is no gold standard diagnosis of a malaria attributable fever; an individual can have malaria parasites in her blood and a fever, but may have developed immunity that allows her to tolerate the parasites and the fever is being caused by another infection. We define the MAFF using the potential outcome framework and show what assumptions underlie current estimation methods. Current estimation methods rely on the parasite density being correctly measured. This generally does not hold because (i) fever kills some parasites and (ii) parasite density has measurement error. In the presence of these problems, we show current estimation methods do not perform well. We propose a novel estimation method based on exponential family g-modeling. Under the assumption that the measurement error mechanism and the magnitude of the fever killing effect are known, we show that our proposed method provides approximately unbiased estimates of the MAFF.

EMAIL: dsmall@wharton.upenn.edu

57b. INVITED ORAL POSTER:

Data-Adaptive Variable Importance with CV-TMLE

Alan E. Hubbard*, University of California, Berkeley
Chris Kennedy, University of California, Berkeley

Big Data holds the promise of discovering relevant relationships using mining algorithms. We propose an approach that uses cross-validation to define the target estimand, estimate the parameter and provide robust inference. We present an estimator (and corresponding central limit theorem) of this data adaptive target parameter: cross-validated targeted

maximum likelihood estimation (CV-TMLE). The algorithm uses the training sample to define levels of an explanatory variable and uses the validation sample to estimate a variable importance comparing these levels; Superlearning and causal-inference inspired estimands are both used to define the target levels and to estimate the adjusted association within the validation sample. We borrow power across the entire sample by a global bias-reduction step (CV-TMLE), which also results in asymptotically-linearly estimators with inference based on the influence curve. Within the proposed algorithm (varImpact), data-cleaning, accounting for missing data, model-fitting, etc., is automated and the algorithm is scaleable for use with large samples and many variables. We apply varImpact to both clinical and epidemiological data.

EMAIL: hubbard@berkeley.edu

57c. The Generalized Front-Door Formula for Identification of Partial Causal Effects

Isabel Fulcher*, Harvard University
Eric Tchetgen Tchetgen, Harvard University
Ilya Shpitser, Johns Hopkins University

Unmeasured confounding is an unfortunate reality of observational studies, making methods robust to this issue invaluable. Pearl's front-door criterion provides conditions for identifying a causal effect of the exposure on the outcome that is robust to unmeasured confounding. Identification of this causal effect requires that (1) a set of observed intermediate variables intercepts all directed paths from the exposure to the outcome and (2) there is no unmeasured confounding of the exposure-intermediate and intermediate-outcome relations. We propose a generalization of the front-door formula by relaxing condition (1) thereby allowing a direct causal effect of the exposure on the outcome. Thus, our formula partially recovers the causal effect of the exposure on the outcome to the extent that it is captured by the observed intermediate factors, despite unmeasured confounding of the exposure-outcome relation and provided assumption (2) holds. We describe parametric and semiparametric doubly robust estimators for this partial causal effect, evaluate the statistical properties of the estimators in a simulation study, and illustrate the methods in several applications.

EMAIL: isabelfulcher@g.harvard.edu

57d. Evaluating Effectiveness of Case-Matching for Exposure-Response Analysis

Jayson D. Wilbur*, Metrum Research Group
Manish Gupta, Bristol-Myers Squibb
Chaitali Passey, Bristol-Myers Squibb
Amit Roy, Bristol-Myers Squibb

Characterization of exposure-response relationships can be challenging in the presence of confounding factors that affect both pharmacokinetic properties as well as response. In such situations, virtual randomization using case-matching of treatment arm subjects has been proposed to select control arm subjects for inclusion in the analysis. We present two approaches to evaluate the effectiveness of the matching with respect to pharmacokinetic properties. The proposed evaluation methods are illustrated for a two-arm clinical trial, in which treatment arm subjects with exposures in the lowest quartile are matched to control arm subjects by propensity score matching. The effectiveness of the matching is assessed by: (1) holding out half the subjects in the treatment arm and attempting to match within the treatment arm and (2) reverse matching the identified control subjects back to the treatment arm, and comparing exposure to what would be expected. The validity of these methods was assessed for several simulated scenarios and both evaluation methods were found to be useful for assessing the effectiveness of the matching with respect to pharmacokinetic properties.

EMAIL: jaysonw@metrumrg.com

57e. Bayesian Causal Inference using Gaussian Process

Jinzhong Liu*, University of Cincinnati
Bin Huang, Cincinnati Children's Hospital Medical Center
Siva Sivaganesan, University of Cincinnati

Bayesian approach to causal inference has traditionally been modeling the outcome mechanism, ignoring the fact that treatments are selectively assigned. On the other hand, frequentist approach has been relying on removing treatment selection bias via baseline covariates matching or propensity score methods, both of which are two-step methods. This article

proposes a Bayesian approach for estimating the population mean treatment effect using Gaussian process (GP), which accomplishes matching and modeling outcome mechanism in a single step. We demonstrate a close relationship between matching method and GP for estimating average treatment effect. The proposed method uses a distance similar to Mahalanobis distance but determines the range of matching automatically without imposing a caliper arbitrarily. Our simulation study results suggest that GP leads to an accurate and more efficient estimate than the linear regression with adjustment for propensity score, inverse probability weighting and BART. We illustrate the proposed method using data obtained from the PRCOIN for estimating the effect of early aggressive biological disease modifying anti-rheumatic drugs.

EMAIL: ljzy208@gmail.com

57f. Causal Inference on Quantiles with Application to Electronic Health Records

Dandan Xu*, University of Texas, Austin
Michael J. Daniels, University of Texas, Austin
Almut G. Winterstein, University of Florida

We propose methods for causal inference on quantiles using a Bayesian nonparametric (BNP) approach in the presence of many confounders. In particular, we define relevant causal quantities and specify BNP models to avoid bias from restrictive parametric assumptions. We first use Bayesian additive regression trees (BART) to model the propensity score and then construct the distribution of potential outcomes given the propensity score using a Dirichlet process mixture (DPM) of normals model. We thoroughly evaluate the operating characteristics of our approach and compare it to Bayesian and frequentist competitors. We use our approach to answer an important clinical question involving acute kidney injury using electronic health records.

EMAIL: xudandanpuck@gmail.com

57g. Bayesian Latent Mediation Model

Chanmin Kim*, Harvard School of Public Health
Michael J. Daniels, University of Texas, Austin
Yisheng Li, University of Texas MD Anderson Cancer Center

We propose a Bayesian semiparametric method to estimate the cluster-specific direct and indirect effects where the clusters are formed by both individual covariate profiles and individual effects. These cluster-specific direct and indirect effects can be estimated through a set of regression models where the coefficients are clustered by setting a stick-breaking prior on the mixing distribution of the allocating latent variable. To let cluster- ing be appropriately informed by individual direct and indirect effects, an informative prior on the coefficient for the effect is specified. We conduct simulation studies to assess performance of the proposed method compared to other clustering methods. We use this approach to estimate the cluster-specific causal direct and indirect effects of an expressive writing intervention for patients with renal cell carcinoma (Milbury et al., 2014).

EMAIL: ckim@hsph.harvard.edu

57h. Congenial Causal Inference with Binary Structural Nested Mean Models

Linbo Wang*, Harvard University
Thomas S. Richardson, University of Washington
James M. Robins, Harvard University

Structural nested mean models (SNMMs) are among the fundamental tools for inferring causal effects of time dependent exposures from longitudinal studies. With binary outcomes, however, the conventional methods for estimating multiplicative and additive SNMM parameters suffer from the variation dependence between the SNMMs and the baseline risk, which is a nuisance model, whereas with the logistic SNMMs, there is no guarantee for a valid test of the causal null hypothesis even in randomized trials. This weakness has hinged the spread of SNMMs in epidemiological practice, where binary outcomes are commonly present. We tackle this problem by proposing the conditional generalized odds product and confounder blip functions as preferred nuisance models for the multiplicative SNMM. Our novel nuisance models are variation independent of the multiplicative SNMMs and hence allow congenial causal inference. Maximum likelihood may be used for parameter estimation without resorting to semiparametric theory. In addition, we point out that the additive SNMMs are not congenial of each other with

binary outcomes. Our approach is illustrated via simulations and a data analysis.

EMAIL: linbowang@g.harvard.edu

57i. Matching Estimators for Causal Effects of Multiple Treatments

Anthony D. Scotina*, Brown University
Roee Gutman, Brown University

Matching estimators for average treatment effects are widely used in the binary treatment setting, in which missing potential outcomes are imputed as the average of observed outcomes of all matches for each unit. With more than two treatment groups, however, estimation using matching requires additional techniques. In this paper, we propose a nearest-neighbors matching estimator for use with multiple, nominal treatments, and use simulations to show that this method is efficient and has coverage levels that are close to nominal.

EMAIL: Anthony_Scotina@brown.edu

57j. Causal Moderation Assessment with Zero-Inflated Data

Robert Gallop*, West Chester University

Within the NIDA Clinical Trials Network (CTN) studies CTN0018 and CTN0019, two primary measures: drug use and sex under the influence, are count responses with a large amount of zeros. Aim of the CTN studies focused on greater effectiveness for the five session intervention (experimental treatment), relative to the one session intervention (standard treatment). Mediation models assess mechanism through which the interventions work. Moderation models assess modifiers of the magnitude of the intervention effect. Standard approaches to mediation and moderation assumes the mediators and moderators are not confounded by unobserved variables. This assumption is not reasonable for any post randomization variable. Hence, the intervention, mediation, and moderation effects, may not be able to be distinguished from the effects of confounders. Our goal is to present an application of causal inference methods for zero-inflated measures in the assessment of mediation and moderation. The combined CTN data provides sufficient power for this investigation. Comparison between the standard methods and causal method for zero-inflated data will be made.

EMAIL: rgallop@wcupa.edu

57k. Dimension Reduction Methods for Continuous Treatment Regimes in Causal Inference

Samantha Noreen*, Emory University
Qi Long, University of Pennsylvania

Sufficient dimension reduction methods rely heavily on the assumption of conditional independence, as does causal inference modeling via ignorability. Dimension reduction methods such as sliced inverse regression (SIR) and partial least squares (PLS) allow for valid causal inference with reduced covariates when considering binary treatment (Ghosh, 2011). The ideas underlying these concepts of dimension reduction and causal estimands of interest can be extended to general treatment regimes. The use of SIR and PLS methods for dimension reduction result in reduced covariates as balancing scores to subsequently provide consistent estimates for the causal effect of interest, while taking advantage of traditional propensity score methods such as inverse probability weighting.

EMAIL: snoreen@emory.edu

58. Nonparametric Methods

› Adjusted Empirical Likelihood-Based Inference for Partially Linear Models

Haiyan Su*, Montclair State University

To improve the accuracy of the normal approximation-based confidence intervals, we consider statistical inference for partially linear models using the adjusted empirical likelihood method. Adjusted empirical likelihood method is an improvement of the empirical likelihood by adding a point to the profile likelihood. It has been shown to preserve all the asymptotic properties of the EL method. The coverage probability of the confidence interval from AEL is also improved particularly when the sample size is small. The test statistics

we developed is shown to be asymptotically chi-square distributed. The numerical performance of the proposed method will be evaluated from numerical simulations and a real study.

EMAIL: suh@mail.montclair.edu

› Monotone Variance Function Estimation using Quadratic Splines

Soo Young Kim*, Colorado State University
Haonan Wang, Colorado State University
Mary C. Meyer, Colorado State University

Various methods for estimating variance functions in heterosecedastic regression models have been developed over the years. We consider the problem of estimating a monotone variance function using quadratic splines. The method is based on the maximum likelihood principle, and its computation is carried out through the convex programming. The convergence rate of the monotone spline approximant is derived, and the same optimal rate is preserved as an unconstrained spline approximant. Simulation results show that our proposed method is comparable or outperforms existing methods under various settings. In addition, the application of the method is illustrated through the analysis of several real datasets.

EMAIL: soo.kim@colostate.edu

› Kernel Density Estimation based on Progressive Type-II Censoring

Hani Samawi*, Georgia Southern University
Amal Helu, Carnegie Mellon University, Qatar
Haresh Rochani, Georgia Southern University
Jingjing Yin, Georgia Southern University
Robert Vogel, Georgia Southern University

Progressive censoring is essential for researchers in industry as a mean to remove subjects before the final termination point. Recently, kernel density estimation has been intensively investigated due to its nice properties and applications. In this paper we investigate the asymptotic properties of the kernel density estimators based on progressive type-II censoring and their application to hazard function estimation. A bias adjusted kernel density estimator is also suggested. Our simulation indicates that the kernel density estimates under progressive type-II censoring is competitive with kernel density estimates under simple random sampling.

EMAIL: samawi.hani2@gmail.com

59. Cancer Genomics

› A Novel Statistical Approach for Identification of the Master Regulator Transcription Factor

Sinjini Sikdar*, University of Florida
Susmita Datta, University of Florida

Transcription factors play key roles in carcinogenesis and are gaining popularity as potential therapeutic targets in drug development. A master regulator transcription factor often appears to control most of the regulatory activities of other transcription factors and genes and is at the top of the hierarchy of transcriptomic regulation. It is important to identify this master regulator for proper understanding of the associated disease process. We present a two-step method for identification of master regulator, where at the first step we test for the existence of master regulator in the system. At the second step, we identify the master regulator, if there exists one. In simulations, our method performs reasonably well in validating the existence of master regulator. In application to real datasets, our method identifies meaningful master regulators. Understanding the regulatory structure and finding the master regulator help narrowing the search space for identifying biomarkers for complex diseases. Also our method provides an overview of the regulatory structure of the transcription factors which control the global gene expression profiles and cell functioning.

EMAIL: sinjinisikdar@ufl.edu

› Statistical Methods to Associate Intra-Tumor Heterogeneity with Clinical Outcomes

Paul Little*, University of North Carolina, Chapel Hill
Wei Sun, Fred Hutchinson Cancer Research Center
Danyu Lin, University of North Carolina, Chapel Hill

Many methods have been developed to infer intra-tumor heterogeneity using whole genome or exome-seq data collected from

tumor samples. However, there are few methods to associate clinical outcomes with the degree of intra-tumor heterogeneity. One challenge for such association analysis is to account for the uncertainty in estimating intra-tumor heterogeneity. We have developed a new method to address this challenge. We summarize the degree of intra-tumor heterogeneity by an entropy measurement and then associate it with clinical outcomes such as survival time. Our method allows one to model the read count data and clinical outcome in a joint likelihood framework, with entropy as a latent variable. We can also simplify our method as a two-step approach to first estimate entropy and then perform association analysis. We will compare different approaches in simulation and real data analyses.

EMAIL: pllittle@email.unc.edu

› Utilizing Patient-Level Characteristics for Identification of Cancer Driver Genes

Ho-Hsiang Wu*, National Cancer Institute, National Institutes of Health
Nilanjan Chatterjee, Johns Hopkins Bloomberg School of Public Health
Bin Zhu, National Cancer Institute, National Institutes of Health

In cancer genomic studies, identifying driver genes associated with cancer initiation and progression is crucial for diagnosis and treatment. For this aim, a probability model is established and a modified one-sided score test is proposed to discover the driver gene with higher frequency of non-silent mutations than its background rate. Distinct from other existing methods assuming homogeneous selection pressures across subjects, the proposed method considers subject-level risk factors and is able to detect selection pressure heterogeneity among subjects. Applied to the lung squamous cell carcinoma of The Cancer Genome Atlas (TCGA) dataset, the proposed method is shown to properly control the false positive rate, identify known cancer driver genes, and discover new ones characterized by the subject-level risk factors.

EMAIL: hohsiang84@gmail.com

› Estimate Penetrance of Breast and Ovarian Cancer in Asian Women Carrying BRCA1/2 Mutation

Lingjiao Zhang*, University of Pennsylvania
Xinglei Chai, University of Pennsylvania
Timothy R. Rebbeck, Dana-Farber Cancer Institute
Jinbo Chen, University of Pennsylvania
Ava Kwong, University of Hong Kong

It is well known that germline mutations in BRCA1/2 genes increase the risk of breast cancer and are associated with an early age at onset. The penetrance of breast and ovarian cancer in BRCA1/2 mutation carriers has been well characterized for Caucasian women, but only limited data exists for Asian women for whom the breast cancer incidence and mutation prevalence are different. We estimated the penetrance of breast cancer in Asian women carrying BRCA1/2 mutation using data collected from Hong Kong Hereditary Breast Cancer Family Registry. We adopted the kin-cohort design that treated the first-degree relatives of the probands as an ascertained cohort, and modified an existing marginal likelihood approach to obtain the age-specific estimates. Our modification was more efficient by incorporating the available genotype data of the first-degree relatives. We also conducted a meta-analysis on available estimates of breast cancer penetrance in BRCA1/2 mutation carriers in Asian women which may be useful for counseling and decision making.

EMAIL: lingjiao@mail.med.upenn.edu

› Pathway-Guided Integrative Analysis of High Throughput Genomic Datasets to Improve Cancer Subtype Identification

Dongjun Chung*, Medical University of South Carolina
Linda Kelemen, Medical University of South Carolina
Brian Neelon, Medical University of South Carolina

The accurate and robust identification of cancer subtypes using high throughput genomic data is of critical importance, because it will inform molecularly-based tumor classification and shared pathogeneses, which can be used to develop more effective prevention and intervention strategies to reduce the burdens of patients suffered from cancers. However, it still remains a challenging task 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 statistical approach for the simultaneous identification of cancer subtypes and driver molecular features by integrating multiple types of cancer genomics datasets with biological pathway information. I will discuss the power of the proposed method with simulation studies and application to TCGA datasets.

EMAIL: chungd@musc.edu

› Comparison of Correlated-Data Methods for Family-Based Genetic Studies

Ana F. Best*, National Cancer Institute, National Institutes of Health
Sharon A. Savage, National Cancer Institute, National Institutes of Health
Philip S. Rosenberg, National Cancer Institute, National Institutes of Health

A common study design in clinical genetics involves data collection on affected mutation carriers and their unaffected family members. These data are often analyzed using standard clustering methods, grouping subjects at the family level. However, this does not account for variability in between-subject correlation within each family. These correlations, due to shared genetic and environmental factors, are highest in closely-related family members but reduced for distantly-related subjects. Mixed effects and GEE methods can be used to explicitly incorporate family structure into inference, but they require additional effort to code pedigrees. Our goal is to empirically assess the inferential benefits of using these methods, variance correction and bias reduction, compared with naïve analysis or standard clustering. We compare methods through simulations and using data from two family-based studies of Li-Fraumeni syndrome (LFS): quantification of first- and second- cancer incidence and survival in a North American LFS cohort (survival analysis), and comparison of telomere attrition rates between North American and Brazilian LFS cohorts (linear and logistic regression).

EMAIL: ana.best@nih.gov

› Tumor Purity Improves Cancer Subtype Classification from DNA Methylation Data

Hao Feng*#, Emory University
Weiwei Zhang, Shanghai Normal University
Hao Wu, Emory University
Xiaoqi Zheng, Shanghai Normal University

Tumor sample classification has long been an important task in cancer research. Classifying tumor into different subtypes greatly benefits therapeutic development and facilitates application of precision medicine on patients. In practice, solid tumor tissue sample obtained from clinical settings is always a mixture of cancer and normal cells. Thus, the data obtained from these samples are mixed signals. The "tumor purity", or the percentage of cancer cells in cancer tissue sample, will bias the clustering results if not properly accounted for. In this paper, we developed a model-based clustering method and software package InfiniumClust, which uses DNA methylation microarray data to infer tumor subtypes with the consideration of tumor purity. Simulation studies and the analyses of The Cancer Genome Atlas (TCGA) data demonstrate improved results compared with existing methods.

EMAIL: hfeng5@emory.edu

60. Measurement Error

› There is No Impact of Exposure Measurement Error on Latency Estimation in Linear Models in Many Cases

Sarah B. Peskoe*, Harvard School of Public Health
Donna Spiegelman, Harvard School of Public Health
Molin Wang, Harvard School of Public Health

Identification of the latency period of a time-varying exposure is key when assessing many environmental, nutritional, and behavioral risk factors. A pre-specified exposure metric involving an unknown latency parameter is often used in the statistical model for the exposure-disease relationship. Likelihood-based methods have been developed to estimate this latency parameter for generalized linear models, but are nonexistent for scenarios where the exposure is measured with error, as is usually the case. Here, we explore the performance of naive estimators for both the latency parameter

and the regression coefficients, which ignore exposure measurement error, assuming a linear measurement error model. We prove that in many scenarios under this general measurement error setting, the least squares estimator for the latency parameter remains consistent, while the regression coefficient estimates are inconsistent as previously obtained under standard measurement error models where the primary disease model does not involve a latency parameter. The findings are illustrated in a study of body mass index in relation to physical activity in the Health Professionals Follow-up Study.

EMAIL: speskoe@fas.harvard.edu

› A Bayesian Approach for Correcting Exposure Misclassification in Meta-Analysis

Qinshu Lian*, University of Minnesota
Haitao Chu, University of Minnesota

In the epidemiological studies, misclassification of the exposure is common. Internal data and external data are the two common validation sources used to account for misclassification in a single observational study. The former one is recognized as a rigorous source, while the latter one is only used in sensitivity analyses in general due to its strong assumptions. In a meta-analysis, evidence from different studies is synthesized. It is imperative to properly account for misclassification to obtain valid estimates. However, it is rare that the internal validation is conducted in every included study. In this paper, we propose a novel framework for correcting misclassification in a meta-analysis through a Bayesian likelihood-based approach. We extend the current scope of using external validation data by relaxing the assumptions through mixed effects models. We simultaneously synthesize main studies with several validation sources, including partial internal validation data, external validation studies and expert opinions. Our model is evaluated through both synthetic data and real data.

EMAIL: lianx025@umn.edu

› Considerations for Analysis of Time-to-Event Outcomes Measured with Error: Bias and Correction with SIMEX

Eric J. Oh*, University of Pennsylvania
Bryan Shepherd, Vanderbilt University
Pamela A. Shaw, University of Pennsylvania

For time-to-event outcomes, a rich literature exists on the bias introduced by covariate measurement error in regression models such as the Cox model, and the methods of analysis to address this bias. By comparison, less attention has been given to understanding the impact or addressing errors in the failure time outcome. For many diseases, the timing of an event of interest (such as time to AIDS progression) can be difficult to assess or reliant on self-report and therefore error-prone. With non-linear outcomes, even non-differential error can introduce bias into estimated associations of interest. We compare the performance of two common models, the Cox model and the Weibull regression model in the setting of measurement error in the failure time outcome. We introduce an extension of the SIMEX method to correct for bias in hazard ratio estimates from the Cox model and discuss other analysis options to address measurement error in the response. Detailed numerical studies are presented to examine the performance of SIMEX under varying levels and parametric forms of the error in the outcome. We further illustrate the method with observational data on HIV outcomes.

EMAIL: ericoh@mail.med.upenn.edu

› Semiparametric Analysis of the Linear Transformation Model for Interval Censored Data with Covariate Measurement Error

Soutrik Mandal*, Texas A&M University
Suojin Wang, Texas A&M University
Samiran Sinha, Texas A&M University

We propose a consistent method for analyzing linear transformation models when the observed data are subject to interval censoring and covariates are measured with errors. The linear transformation model includes the proportional hazards model and the proportional odds model as special cases. In this context, maximum likelihood method is complex, and

may not be robust to some model assumptions. Therefore, we propose to estimate the model parameters by solving a set of unbiased estimating equations while for handling interval censoring and covariate measurement errors we develop a multiple imputation approach. Specifically, for handling measurement errors in covariates, we use a nonparametric approach by adopting a density deconvolution technique. Operating characteristics of the proposed approach are judged via simulation studies, and we apply the proposed method to analyze a real data set from an AIDS clinical trial.

EMAIL: smandal@stat.tamu.edu

› Causal Mediation Analysis with Survival Data: Interaction and Measurement Error

Ying Yan, University of Calgary
Lingzhu Shen*, University of Calgary
Gemai Chen, University of Calgary

Causal mediation analysis is a useful tool to examine how an exposure variable causally affects an outcome variable through an intermediate variable (mediator). In recent years, there has been increasing research interest in mediation analysis with survival data. The existing literature requires accurate measurements of the mediator and the confounder, which could be infeasible in biomedical studies. Furthermore, the current identification results of causal effects under the additive hazards model do not allow for exposure-mediator interaction, which may be unappealing in mediation analysis. In this talk, we derive identification results of causal effects under the additive hazards model with exposure-mediator interaction. Furthermore, we provide bias formulas to quantify the impact of mismeasuring the mediator or the confounder (or both) on causal effects estimation, and propose consistent measurement error correction methods in the absence/presence of exposure-mediator interaction. The performance of the proposed methods is demonstrated in simulation studies and an AIDS study. This is a joint work with Dr. Ying Yan and Dr. Gemai Chen at the University of Calgary, Canada.

EMAIL: lingzhu.shen@ucalgary.ca

› Correcting for Misclassification in the Partial Population Attributable Risk

Benedict Wong*, Harvard University
Donna Spiegelman, Harvard School of Public Health

Estimation of the partial population attributable risk (pPAR) has become an important goal in public health research, because it describes the proportion of disease cases that could be prevented if an exposure 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. We assume that the relationship between the binary disease status and the true covariates follows a logistic regression model, and we also use several methods to model the misclassification process, including the matrix method, inverse matrix method, and a series of logistic regression models. We present a method for obtaining point and interval estimates of the pPAR in the presence of misclassification, where we maximize the full data likelihood, to obtain maximum likelihood estimates for the parameters of the logistic regression models for the disease and for the misclassification process. We apply this to data from the Health Professionals Follow-Up Study.

EMAIL: wong01@fas.harvard.edu

61. Missing Data Methods

› Statistical Approaches that Account for Accelerometer Non-Wear and the Missing Mechanism at the Day-Segment, Day, and Participant Levels

Rebbecca Wilson*, University of North Carolina, Chapel Hill
Daniela Sotres-Alvarez, University of North Carolina, Chapel Hill
Kelly R. Evenson, University of North Carolina, Chapel Hill
Shrikant I. Bangdiwala, University of North Carolina, Chapel Hill

Accelerometers provide objective measures of physical activity and sedentary behavior when the device is worn for one week during all waking hours. A challenge in most free-living accelerometer studies is accounting for time when the accelerometer is not worn. Non-wear, defined by a period of consecutive zero counts, is treated as missing data. Standard practice summa-

rizes accelerometer data at the day level (e.g., minutes/day in sedentary), and then averages across days if there is at least a minimum number of days and a minimum wear time per day. This approach assumes physical activity and sedentary behavior are MCAR during non-wear time, but physical activity and sedentary behavior are related to non-wear time. The aim of this study was to assess the impact in the bias and precision of mean estimates and percent below a threshold under different missing data mechanisms (MCAR, MAR, NMAR). We simulated complete data under different scenarios and then created several different missing mechanisms scenarios. We implemented multiple imputation and the generalized linear mixed model, accounting for missing data at the day-segment, day, and participant levels.

EMAIL: dsotres@unc.edu

› Privacy-Preserving Methods for Horizontally Partitioned Incomplete Data

Yi Deng*, Emory University
Qi Long, Emory University

Distributed health data networks that leverage data from multiple sources have drawn substantial interests in recent years. However, such networks face challenges in the presence of missing data. The current state-of-the-art missing data methods require pooling data into a central repository before analysis, which may not be practical. In this paper, we propose a privacy-preserving framework, in which each institution with a particular data source utilizes the local private data to calculate only necessary intermediate statistics which are then shared across all institutions for the proposed distributed algorithms. As such, the proposed framework as well as the involved distributed algorithms are privacy-preserving since no individual-level data are shared, leading to lower hurdles for collaboration across multiple institutions and stronger public trust with more institutions participating. To evaluate our proposed methods, we conduct simulation studies and then show the proposed privacy-preserving methods perform as well as the methods using the pooled data.

EMAIL: ydeng26@emory.edu

› An Extended Propensity Score Approach to Account for Missing Confounders when Estimating Causal Effects

Katherine Evans*#, Harvard University
Eric Tchetgen Tchetgen, Harvard University
Francesca Dominici, Harvard University

The effect of treatment on the treated (ETT) is a common parameter of interest in causal inference. Identification of ETT typically relies on an assumption of no unobserved confounding. When information on a subset of confounders is not observed in a main study, data from a validation study with more detailed confounders may be used to help mitigate confounding. Such methods have been extended to address missing confounder data in a main-validation study context; however existing methods rely on restrictive assumptions that are unlikely to hold in practice. To address this problem, we develop a novel approach which entails constructing an extended propensity score which preserves essential properties of a standard propensity score, but with the additional advantage that it can be evaluated even for subjects with missing confounders. The finite sample performance of the proposed approach is evaluated and compared to existing methods in simulation. The proposed approach is also illustrated in an application examining the effect of surgical resection on survival time among Medicare beneficiaries with malignant neoplasm of the brain using SEER-Medicare for the validation study.

EMAIL: kevans@fas.harvard.edu

› Bayesian Single Index Model with Covariates Missing at Random

Kumaresh Dhara*, Florida State University
Debdeep Pati, Florida State University
Debajyoti Sinha, Florida State University

Bayesian single index model is a highly promising dimension reduction tool for an interpretable modeling of the non-linear relationship between the response and its predictors. However, existing Bayesian tools in this area suffer from slow mixing of the Markov Chain Monte Carlo (MCMC) computational tool and also lack the ability to deal with missing covariates. To circumvent these practical problems, we present a new

Bayesian Single Index model with MCMC algorithm using a mode-alignment based proposal density for the index vector for an efficient Metropolis Hastings (MH) algorithm to sample from full conditional distribution. Our method leads to an interpretable model and inference, the efficient evaluation of the likelihood, fast convergence of the MCMC, and a first time extension of inference to missing at random covariates. We also give theoretical justifications of our method by proving posterior convergence.

EMAIL: k.dhara@stat.fsu.edu

› A Robust Association Test in the Presence of Missing Data with Applications to Integrative Genomics Analysis

Kin Yau Wong*, University of North Carolina, Chapel Hill
Donglin Zeng, University of North Carolina, Chapel Hill
Danyu Lin, University of North Carolina, Chapel Hill

Genome-wide association studies assess the association between single nucleotide polymorphisms (SNPs) and a disease by investigating the genetic variants separately. The mechanism by which the SNPs affect the disease is not taken into account. To improve power, tests that consider SNPs along with other genomics data types such as mRNA expression data have been proposed. However, the joint tests require complete data of all data types, which may not be available in reality. In this study, we propose to impute the missing genomics data using other observed genomics data and covariates based on a semiparametric model. We show that while naive imputation methods may result in a biased score statistic, score statistic obtained from the proposed semiparametric imputation model is always unbiased. The proposed approach is applicable to continuous, binary, and survival outcome variable. We show by simulation studies that the proposed method has correct type I error and decent power under different distributions of the missing variable. We apply the proposed method to a data set from The Cancer Genome Atlas.

EMAIL: alexwky@live.unc.edu

› A High-Dimensional Multivariate Selection Model for Proteomics Data with Batch-Level Missingness

Jiebiao Wang*, University of Chicago
Pei Wang, Icahn Medical School at Mount Sinai
Donald Hedeker, University of Chicago
Lin Chen, University of Chicago

In quantitative proteomics, mass tag labelling techniques, such as iTRAQ, have been widely adopted in mass spectrometry experiments. These techniques allow peptides/proteins from multiple samples of a batch to be quantified in a single experiment, but they come at a cost of severe batch effects and batch-level non-ignorable missing data. Taking into account batch effects and missing data, we develop a multivariate selection model to jointly analyze multiple peptides of each protein. To facilitate the computation for high-dimensional outcomes, we employ a log link in the missing-data model and introduce an item response theory -type random effect structure that reduces the number of variance components to the dimension of outcomes. An EM algorithm is used for estimation. Simulations show the advantages of the proposed method in reducing estimation bias, controlling type I error rates and improving power, compared to conventional methods. We apply the proposed method to an iTRAQ-based proteomics dataset and identify proteins related to breast cancer. The proposed method can be applied to general multivariate analyses based on clustered data with outcome-dependent missingness.

EMAIL: jwang88@uchicago.edu

62. Translational Research/Science

› Characterizing Protein Backbone Geometry through Statistical Analyses: A Comparative Study

Saptarshi Chakraborty*, University of Florida
Samuel W.K. Wong, University of Florida

The backbone geometry of a protein plays an important role in the understanding and prediction of protein structure in bio-informatics. This geometry is summarized in the Ramachandran plot, a scatterplot of dihedral or torsion angles. Various non-parametric and model-based parametric

approaches have been proposed in the literature to analyze such data. While a non-parametric approach has a wide range of applicability, one major problem is that the associated model usually has a large number of parameters and a very high degree of complexity. The fitted model surface can be extremely rough for moderate sample sizes, which is unsuitable for predictive applications and difficult to interpret. In contrast, a single parametric density, though simple and easy to interpret, is usually not adequate to model complex data. Finite mixture models of these densities are an appealing alternative; however, fitting mixture models to bivariate angular data is not straightforward and no optimal strategy is suggested in the literature. In this talk we examine the methods available for analyzing data on protein dihedral angles and make a comparative discussion.

EMAIL: c7rishi@ufl.edu

› Generalized Factor Analysis for Integrative Analysis with Incorporation of Structural Information

Changgee Chang*, University of Pennsylvania
Yize Zhao, Cornell University
Qi Long, Emory University

Various omics data are massively generated in many biomedical and clinical studies, and it is desirable to pool such data in order to improve the power of identifying important molecular signatures, which are often subtle in the individual data sets. However, such integrative data analyses entail new analytical and computational challenges. To address these challenges, we propose a latent factor based Bayesian method. The Bayesian sparse generalized factor analysis (GFA) features integrating multiple omics modalities with incorporation of biological knowledge through the use of a mixture of spike and slab prior and Markov random field prior. The methods can also accommodate both continuous and discrete data and are expected to achieve improved feature selection and prediction in identifying disease subtypes and latent drivers, compared to existing methods. Efficient EM algorithms are presented. Simulation study confirms the superiority of the proposed method. We applies the method to a real world application and obtains biologically meaningful findings.

EMAIL: changgee@mail.med.upenn.edu

› Inverse Probability Weighting with a Response-Informed Calibrated Propensity Score: A Robust and Efficient Approach for Treatment Effects in Large Observational Studies

David Cheng*, Harvard School of Public Health
Abhishek Chakrabortty, University of Pennsylvania
Ashwin Ananthakrishnan, Massachusetts General Hospital
Tianxi Cai, Harvard School of Public Health

Adjusting for the propensity score (PS) is a common approach to estimate treatment effects in observational studies (OS). Two widely used estimators include the inverse probability weighting (IPW) and doubly-robust (DR) estimators. The performance of both deteriorate when underlying parametric models are mis-specified and when adjusting for a large number of covariates. We propose a response-informed calibrated PS estimator (RiCaPS) that uses nonparametric smoothing to calibrate an initial parametric PS while incorporating information from covariates conditionally associated with the response. Regularization is used for the parametric models to accommodate a large number of covariates. We show that an alternative IPW estimator based on this calibrated PS has the double-robustness and local semiparametric efficiency properties. The RiCaPS IPW enjoys additional robustness and efficiency benefits compared to standard methods under model mis-specification and when adjusting for many covariates. Simulations confirm these favorable properties in finite samples. We illustrate the method in an electronic medical records (EMR) study and a cohort study.

EMAIL: dcheng01@fas.harvard.edu

› Selection of Effective Scores for Treatment Effect Estimation

Ying Zhang*, University of Wisconsin, Madison
Lei Wang, University of Wisconsin, Madison
Menggang Yu, University of Wisconsin, Madison
Jun Shao, University of Wisconsin, Madison

Under the assumption of treatment and potential outcomes are independent conditional on all covariates, valid treatment effect estimators can be obtained using nonparametric inverse propensity weighting and/or regression, which are popular because

no model on propensity or regression is imposed. To obtain efficient treatment effect estimators, typically the set of all covariates can be replaced by lower dimensional sets containing linear combinations of covariates, which are called effective scores. We propose to construct an effective score separately for each treatment and show that the resulting asymptotic variance of treatment effect estimator reaches a lower bound that is smaller than those based on other effective scores. Since the effective scores have to be estimated, for example, using sufficient dimension reduction, we derive theoretical results on when the efficiency of treatment effect estimation is affected by estimating effective scores. Our theory is complemented by some simulation results and an illustration is also made using data from an Accountable Care Organization study.

EMAIL: yyzhang818@gmail.com

› Sequential Rank Agreement Methods for Comparison of Ranked Lists

Claus T. Ekstrom*, University of Copenhagen
Thomas A. Gerds, University of Copenhagen
Andreas K. Jensen, University of Copenhagen

Ranked lists of predictors are common occurrences as part of statistical analyses and they frequently appear in the results section of scientific publications. Often, the aim is to find a consensus set of shared predictors from the top of these ranked lists. For example, in high-dimensional data situations like genomics, genes may be ranked differently based on different ranking methods or they may be ranked slightly diffently under varying conditions. We will show how sequential rank agreement can be used to gauge the similarity among ranked lists such that it is possible to make inference - both analytical and simulation-based - about how far the lists agree on the ranking. Our method provides both an intuitive interpretation and can be applied to any number of lists even if these are partially censored. To demonstrate the performance of our approach we show results from a both a simulation study and an application to genomics data, and we illustrate how sequential rank agreement can combine and improve results from both parametric and non-parametric statistical learning algorithms.

EMAIL: ekstrom@sund.ku.dk

63. Causal Inference for Continuous-Time Processes: New Developments

› Mimicking Counterfactual Outcomes to Estimate Causal Effects

Judith J. Lok*, Harvard School of Public Health

In observational studies, treatment may be adapted to covariates at several times without a fixed protocol, in continuous time. Treatment influences covariates, which influence treatment, which influences covariates, and so on. Then even time-dependent Cox-models cannot be used to estimate the net treatment effect. Structural nested models have been applied in this setting. Structural nested models are based on counterfactuals: the outcome a person would have had had treatment been withheld after a certain time. Previous work on continuous-time structural nested models assumes that counterfactuals depend deterministically on observed data, while conjecturing that this assumption can be relaxed. This presentation shows that one can mimic counterfactuals by constructing random variables, solutions to a differential equation, that have the same distribution as the counterfactuals, even given past observed data. These "mimicking" variables can be used to estimate the parameters of structural nested models without assuming the treatment effect to be deterministic.

EMAIL: jlok@hsph.harvard.edu

› Estimation of Mean Outcome on Treatment with Treatment Subject to Informative Discontinuation

Brent A. Johnson*, University of Rochester

Authors have investigated the challenges of statistical analyses and inference in the presence of early treatment termination, including a loss of efficiency in randomized controlled trials and a connection to dynamic regimes in observational studies. Popular estimation strategies for causal estimands in dynamic regimes lend themselves to studies where treatment is assigned at a finite number of points and the extension to continuous treatment assignment is non-trivial. We re-examine this from a different perspective and propose a new estimator for the mean outcome of a target treatment length policy that does not involve

a treatment model. Because this strategy avoids modeling the treatment assignment mechanism, the estimator works for both discrete and continuous treatment length data and eschews bias and imprecision that arise as a result of coarsening continuous time data into intervals. We exemplify the new estimator through numerical studies and the analysis of two data sets.

EMAIL: brent_johnson@urmc.rochester.edu

› Optimizing the Personalized Timing for Treatment Initiation with Random Decision Points

Yebin Tao, University of Michigan
Lu Wang*, University of Michigan
Haoda Fu, Eli Lilly and Company

An important but challenging problem for chronic disease with progressions is to find the optimal personalized timing to initiate a treatment for the next stage of disease condition, given a patient's specific characteristics. We aim to identify the optimal dynamic treatment regime (DTR) amongst a set of regimes predefined by key biomarkers indicating disease severity, which are monitored during a follow-up period. Instead of considering multiple fixed decision stages as in most DTR literature, our study undertakes the task of dealing with continuous random decision points for treatment initiation based on patients' biomarker and treatment history. Under each candidate DTR, we employ a flexible survival model with splines of time-varying covariates to estimate the patient-specific probability of adherence to the regime, and construct an inverse probability weighted estimator for the counterfactual mean utility to assess the DTR. We conduct simulations to demonstrate the performance of our method and further illustrate the application process with an example of insulin therapy initiation among type 2 diabetic patients.

EMAIL: luwang@umich.edu

› Comments on Structural Nested Models in Continuous Time

James M. Robins*, Harvard School of Public Health

I discuss recent developments in continuous time structural nested models

EMAIL: robinsjami1@gmail.com

64. Statistical Methods for Estimating Population Treatment Effects from Trials and Non-Experimental Studies

› Generalizing Study Results: A Potential Outcomes Perspective

Catherine R. Lesko*, Johns Hopkins Bloomberg School of Public Health
Ashley L. Buchanan, University of Rhode Island
Daniel Westreich, University of North Carolina, Chapel Hill
Jessie K. Edwards, University of North Carolina, Chapel Hill
Michael J. Hudgens, University of North Carolina, Chapel Hill
Stephen R. Cole, University of North Carolina, Chapel Hill

When the study sample is not a random sample of the target population, the sample average treatment effect, even if internally valid, cannot usually be expected to equal the average treatment effect in the target population. The utility of an effect estimate for planning and decision making will depend on the degree of departure from the causal effect in the target population due to problems with both internal andexternal validity. Herein, we review concepts from recent literature on generalizability, one facet of external validity, using the potential outcomes framework. Identification conditions sufficient for external validity include: conditional exchangeability; positivity; the same distributions of treatment versions; no interference; and no measurement error. We also require correct model specification. Under these conditions, we discuss how a version of direct standardization (the g-formula, adjustment formula, or transport formula) or inverse probability weighting can be used to generalize a causal effect from a study sample to a well-defined target population, and demonstrate their application in an illustrative example.

EMAIL: clesko2@jhu.edu

› Using Evidence from Randomized Trials and Observational Studies

Issa J. Dahabreh*, Brown University
Sarah Robertson, Brown University

Drawing causal inferences when evidence is derived from multiple studies, only some of which are randomized, is challenging because of confounding in the observational studies

and lack of representativeness in the randomized trials. We present methods that overcome these problems by modeling the probability of trial participation, the propensity score, and the conditional outcome mean. We show that under certain conditions, different models can be combined so as to attain robustness to model misspecification. We use the example of a singly-randomized preference trial of surgery versus medical therapy for coronary artery disease to illustrate the methods.

EMAIL: issa_dahabreh@brown.edu

› To Weight or not to Weight: Estimating Population Treatment Effects by using Propensity Score Matching with Complex Survey Data

David Lenis*, Johns Hopkins University
Nianbo Dong, University of Missouri
Elizabeth A. Stuart, Johns Hopkins Bloomberg School of Public Health

Many studies aim to estimate population causal effects using data from large surveys, and many of these studies use propensity score methods to make those causal inferences as rigorous as possible given the non-experimental nature of the data. However, only few studies address how to incorporate survey sampling weights and other survey design elements when using propensity score methods, especially propensity score matching. This study provides suggestions about how to handle sampling weights when implementing matching in complex survey data. These suggestions are based on the results from Monte Carlo simulations as well as a data-based application. There are two primary conclusions regarding the application of propensity score methods to estimate population treatment effects using complex survey data: (1) sampling weights must be taken into account in the outcome analysis and (2) the estimator of the treatment effect performs better after a weight transfer is implemented. We find that these conclusions hold for different initial values of balance among the covariates and across a range of simulation settings and model misspecification.

EMAIL: dlenis@jhsph.edu

65. Emerging Statistical Challenges in Neuroimaging

› Statistical Challenges in the Multivariate Pattern Analysis of Neuroimaging Data

Kristin A. Linn*, University of Pennsylvania
Bilwaj Gaonkar, University of California, Los Angeles
Jimit Doshi, University of Pennsylvania
Christos Davatzikos, University of Pennsylvania
Russell Taki Shinohara, University of Pennsylvania

Multivariate pattern analysis (MVPA) in neuroimaging comprises one or more statistical or machine learning models applied within a larger imaging analysis pipeline. The goal of MVPA is often to learn and understand patterns of variation encoded in magnetic resonance images (MRI) of the brain that are associated with neurological disease occurrence, progression, and response to therapy. Every model choice that is made during image processing and analysis can have implications with respect to the results and conclusions of neuroimaging studies. Here, attention is given to two important steps within the MVPA framework: 1) standardization of features prior to model training, and 2) the training of a supervised learning model in the presence of confounding. Specific examples focus on the use of the support vector machine, but the general concepts apply to a large set of models commonly used for MVPA. In both cases, we propose novel methods that lead to improved classifier performance and interpretability, and we illustrate the methods on real neuroimaging data from a study of Alzheimer's disease.

EMAIL: klinn@upenn.edu

› Bayesian Spatial Binary Regression for Label Fusion in Structural Neuroimaging

D. Andrew Brown*, Clemson University
Kristin Linn, University of Pennsylvania
Christopher S. McMahan, Clemson University
Russell Shinohara, University of Pennsylvania

Label fusion is a process for image segementation through which multiple atlases are combined for identifying anatomical structures of interest in a brain image. Most existing

approaches are based on standard image processing techniques, constructing a data generating model and then using EM or a related algorithm to find the maximum a posteriori (MAP) estimate of the true segementation while ignoring the uncertainty associated with such estimates. We propose a fully Bayesian approach in which the atlas labels are incorporated as covariates into a spatial binary regression model. Markov chain Monte Carlo is used to simulate the entire posterior distribution of voxel labels, providing much richer information than would be available from a MAP estimator alone. We apply our proposed approach to both simulated data and to the ADNI data for hippocampus segmentation, an important brain structure in the study of Alzheimer's disease.

EMAIL: ab7@clemson.edu

› A Bayesian General Linear Modeling Approach to Cortical Surface-Based Task fMRI Activation Studies

Amanda Mejia*, Indiana University
Ryan Yue, Baruch College, CUNY
Martin Lindquist, Johns Hopkins University
David Bolin, Chalmers University of Technology
Finn Lindgren, University of Edinburgh
Havard Rue, Norwegian University of Science and Technology

A common approach to task fMRI analysis is the general linear model (Worsley and Friston 1995), in which the expected response under the task paradigm is regressed against the observed fMRI time series at each brain location separately. While computationally convenient, this model ignores spatial correlations in the brain. Newer models have been proposed to incorporate spatial dependence within slices; however, these may fail to capture the complex structure resulting from a highly folded cortical geometry. We instead represent fMRI data as a cortical surface (Fischl et al. 1999), since geodesic distances along the surface are more neurologically relevant than Euclidean distances within the volume. We use Bayesian spatial processes to model baseline and amplitude fields while assuming autoregressive errors in the temporal domain. Further, we propose a novel joint posterior probability map method to identify regions of activation and employ an approximate Bayesian inference tool to reduce computational burden. We facilitate population inference by extending the model to the multi-subject case. The method is validated through a simulation study and a motor task fMRI dataset.

EMAIL: mandy.mejia@gmail.com

› Removing Inter-Subject Technical Variability in Magnetic Resonance Imaging Studies

Jean-Philippe Fortin*, University of Pennsylvania
Elizabeth M. Sweeney, Rice University
John Muschelli, Johns Hopkins University
Ciprian M. Crainiceanu, Johns Hopkins University
Russell T. Shinohara, University of Pennsylvania

Magnetic resonance imaging (MRI) intensities are acquired in arbitrary units, making scans non-comparable across sites and between subjects. Intensity normalization is a first step for the improvement of comparability of the images across subjects. However, we show that unwanted inter-scan variability associated with imaging site and other technical artifacts is still present after standard intensity normalization. We propose RAVEL (Removal of Artificial Voxel Effect by Linear regression), a tool to remove residual technical variability. The unwanted variation component is estimated from a control region obtained from CSF, where intensities are known to be unassociated with disease status. We perform a singular value decomposition (SVD) of the control voxels to estimate and remove factors of unwanted variation. We assess the performance of RAVEL using T1-weighted (T1-w) images from more than 900 subjects from the ADNI database. We show that RAVEL performs best at improving the replicability of the brain regions that are empirically found to be most associated with AD, and that these regions are significantly more present in structures impacted by AD.

EMAIL: fortin946@gmail.com

66. Lifetime Data Science

› Analysis of Matched Pair Survival Data - A New Look at an Old Problem

David Oakes*, University of Rochester

Kartsonaki and Cox (Biometrika, 103, pp. 219-224, 2016) recently gave a succinct account including some new

approaches of the analysis of matched pair survival data in the absence of censoring. We discuss possible modifications of their methods to matched pairs data subject to censoring.

EMAIL: david_oakes@urmc.rochester.edu

› Estimating the Risk of Breast Cancer Recurrence Using Cancer Registry Data

Angela Mariotto*, National Cancer Institute, National Institutes of Health
Xiaoqin Xiong, Information Management System
Fanni Zhang, Information Management System
Ruth Etzioni, Fred Hutchinson Cancer Research Center

Cancer recurrence is a key outcome for measuring the burden of illness and in clinical decision making. Yet information about recurrence is not available in population-based data sources. To address the lack of recurrence data we developed a method that uses disease specific survival from cancer registry data to estimate disease-free survival and the risk of recurrence. The method is based on an illness-death process that assumes a recurrence occurs before a cancer death and mixture cure survival models to estimate the proportion of cancer patients who are cured and the proportion who will eventually die of their cancer. For those patients that will eventually recur we estimate their disease free survival by subtracting from their survival time, estimated using the mixture cure survival model, the time from recurrence to death using a deconvolution method. Survival from recurrence to death is estimated using survival for initially diagnosed distant stage disease and adjustments based on published results. Finally, we apply the method to breast cancer survival data from the Surveillance, Epidemiology and End Results (SEER) registries.

EMAIL: mariotta@mail.nih.gov

› Regression Analysis of Residual Life with a Long-Term Survivor Fraction

Megan Othus, Fred Hutchinson Cancer Research Center
Xinyi Zhang, Fred Hutchinson Cancer Research Center
Chen-Hsin Chen, Academia Sinica, Taiwan
Ying Qing Chen* Fred Hutchinson Cancer Research Center

We consider a semiparametric model for regression analysis of residual life with a long-term survivor or cure fraction. An EM algorithm is developed for estimation. Theoretical asymptotic results, including identifiability of the model, are derived and small sample properties are verified via simulation. We apply the proposed model to an adjuvant therapy trial in advanced melanoma and a trial of first-line therapy for acute myeloid leukemia. The model provides an alternative interpretation of the trial results beyond the protocol-specified log-rank tests and Cox regression models.

EMAIL: yqchen@fhcrc.org

67. Novel Statistical Methods for Dealing with Disease Heterogeneity

› Dealing with Missing Subtypes under Weak Assumptions Using Auxiliary Case Covariates

Daniel Nevo*, Harvard School of Public Health
Reiko Nishihara, Harvard School of Public Health
Shuji Ogino, Brigham and Women's Hospital, Dana-Farber Cancer Institute, Harvard Medical School and Harvard School of Public Health
Molin Wang, Harvard School of Public Health

The analysis of time-to-disease data to assess risk factor associations with different disease subtypes is typically based on the assumption that the probability of a missing subtype is independent of the true unobserved subtype given disease diagnosis time and covariates measured regularly. In practice, however, this missing-at-random assumption does not necessarily hold. Here we are motivated by colorectal cancer subtype analysis, where some tumors are more likely to result in a missing tumor tissue, and hence a missing subtype. This missing pattern apparently depends on observable tumor characteristics such as tumor location and grade. We develop a method to conduct valid analysis when additional auxiliary variables are measured for cases only, under a weaker missing-at-random assumption that involves the auxiliary case covariates. Overlooking these covariates will potentially result in biased estimates and efficiency loss. We illustrate the use of our method in the analysis colorectal cancer data from the

Nurses' Health Study cohort, where, apparently, the traditional missing-at-random assumption fails to hold.

EMAIL: nhdne@channing.harvard.edu

› Efficiency Consideration of the Etiological Heterogeneity Evaluation in Case-Case and Case-Control Studies

Molin Wang*, Harvard School of Public Health
Aya Kuchiba, National Cancer Center, Tokyo
Ran Gao, University of Wisconsin, Madison

A fundamental goal of epidemiologic research is to investigate the relationship between exposures and disease risk. Cases of the disease are often considered a single outcome and assumed to share a common etiology. However, evidence indicates that many human diseases arise and evolve through a range of heterogeneous molecular pathologic processes, influenced by diverse exposures. Both the case only and case-control studies can be used to evaluate whether the association of a potential risk factor with disease varies by disease subtype. An interesting question is whether we really gain anything by including controls when evaluating the etiological heterogeneity between subtypes. In this talk, we will present both theoretical and numerical results for validation and efficiency comparisons of the etiological heterogeneity evaluation in the case-case and case-control studies. We used an epidemiology study of plasma free estradiol and breast cancer risk by tumor ER and PR subtypes as an illustration example.

EMAIL: stmow@channing.harvard.edu

› Etiologic Heterogeneity: Strategies for Distinguishing Cancer Subtypes

Colin B. Begg*, Memorial Sloan Kettering Cancer Center

This talk will focus on the information that can be gleaned about etiologic heterogeneity from studies of double primary malignancies. This resource provides unique insights about etiologic heterogeneity that cannot be obtained from traditional epidemiologic studies such as cohort or case-control studies. The principles of the concept will be illustrated using examples involving different cancer types, and an analysis of tumor factors that distinguish most clearly the etiologic subtypes of melanoma will be presented.

EMAIL: beggc@mskcc.org

› Testing for Genetic Association in the Presence of Structured and Unstructured Disease Subtype Heterogeneity

Haoyu Zhang, Johns Hopkins University
Nilanjan Chatterjee*, Johns Hopkins University
Montserrat Garcia-Closas, National Cancer Institute, National Institutes of Health

As sample size for genome-wide association studies continues to rise, there is unprecedented opportunity for obtaining new insights to genetic architecture of complex diseases. Many diseases like breast cancer are intrinsically heterogeneous consisting of subtypes that could be defined by various pathologic and molecular disease characteristics. In this talk, I will describe methods for conducting genetic association scans for a disease taking into account subtype heterogeneity. I will describe a subset based unstructured analysis and a more model based structured analysis for modeling heterogeneity. For both approaches, simple methods for handling missing data in tumor characteristics will be described. Applications will be illustrated based on analysis of a large GWAS (N> 100,000) of breast cancer with multiple tumor characteristics including ER, PR and HER2 status.

EMAIL: nilanjan@jhu.edu

68. Modern Methods for Estimation and Inference from Heterogeneous Data

› Learning Local Dependence in Ordered Data

Guo Yu, Cornell University
Jacob Bien*, Cornell University

In many applications, data come with a natural ordering. This ordering can often induce local dependence among nearby variables. However, in complex data, the width of this dependence may vary, making simple assumptions such as a constant neighborhood size unrealistic. We propose a framework for learning

this local dependence based on estimating the inverse of the Cholesky factor of the covariance matrix. Penalized maximum likelihood estimation of this matrix yields a simple regression interpretation for local dependence in which variables are predicted by their neighbors. Our proposed method involves solving a convex optimization problem that decomposes into independent subproblems that can be solved efficiently in parallel. Our method yields a sparse, symmetric, positive definite estimator of the precision matrix, encoding a Gaussian graphical model. We derive theoretical results not found in existing methods attaining this structure. Empirical results show our method performing favorably compared to existing methods. We apply our method to genomic data to flexibly model linkage disequilibrium. This is joint work with Guo Yu.

EMAIL: jbien@cornell.edu

› Boosting in the Presence of Outliers Via Non-Convex Classification

Alexander Hanbo Li, University of California, San Diego
Jelena Bradic*, University of California, San Diego

Recent advances in technologies for cheaper and faster data acquisition and storage have led to an explosive growth of data complexity in a variety of research areas such as high-throughput genomics, biomedical imaging, high-energy physics, astronomy and economics. As a result, noise accumulation, experimental variation and data inhomogeneity have become substantial. However, machine learning in such settings is known to pose many statistical challenges and hence calls for new methods and theories. Moreover, the impact of outliers or non-gaussianity is far from obvious. In this talk we propose a new boosting framework called Arch Boost. It is designed for augmenting the existing work such that its corresponding classification algorithms with non-convex losses are significantly more adaptable to unknown data contamination. Along with the Arch Boost framework, a family of non-convex losses are proposed which leads to new robust boosting algorithms, named Adaptive Robust Boosting (ARB).

EMAIL: jbradic@ucsd.edu

› Integrative Association Analysis of Multiple Heterogeneous Data Sources

Gen Li, Columbia University
Irina Gaynanova*, Texas A&M University

The growth of data collection and data sharing led to increased availability of multiple types of data collected on the same set of objects. As an example, RNASeq, miRNA expression and methylation data for the same tumor samples are publicly available through the Cancer Genome Atlas project (TCGA). Due to the scale of the data, as well as its heterogeneity, it is typical to analyze each data type separately. In this work we use an exponential family framework as a building block for integrative association analysis of multiple heterogeneous data sources. By learning the low rank decomposition of the natural parameter matrix, we are able to identify the patterns that are common across the data sources as well as source-specific patterns. In contrast to existing approaches, the method can accommodate multiple data sources and types (such as continuous, count, binary), as well as perform variable selection.

EMAIL: irinag@stat.tamu.edu

› Post-Selection Testing for the Graphical Lasso

Max G'Sell*, Carnegie Mellon University
William Fithian, University of California, Berkeley

The graphical lasso is often used to model the dependence structure between variables. However, inferential questions about the resulting solution have traditionally been difficult to answer. We discuss two inferential problems in this setting: testing the significance of selected edges and testing the goodness-of-fit of entire selected models. The first problem sheds some light on the reliability of the specific edges in the graphical lasso solution, while the second has connections to sequential multiple testing and model selection.

EMAIL: mgsell@cmu.edu

69. Modern Survival Analysis in Observational Studies

The Utility of Tracing Studies in Cohorts with Loss to Follow-Up

Richard J. Cook*, University of Waterloo
Nathalie Moon, University of Waterloo
Leilei Zeng, University of Waterloo

Tertiary care facilities often maintain registries of patients with longitudinal follow-up to create a platform for research on the course of chronic disease and the identification of associated risk factors. The observation process in such settings is complex since individuals may miss scheduled visits or make unscheduled visits due to a worsening of symptoms. Long gaps from the last clinic visit to an administrative censoring date may also arise due to patient withdrawal or death, and without further investigation it may not be clear if and when either of these events occurred. We consider design issues for tracing studies in which, subject to budgetary constraints, efforts are made to contact individuals who have not attended the clinic for an undue length of time. The feasibility of efficient selection of individuals for tracing is discussed when interest lies in fitting multistate models to data from a panel observation scheme. Data from a rheumatology clinic are used for motivation and illustration. This is joint work with Nathalie Moon and Leilei Zeng.

EMAIL: rjcook@uwaterloo.ca

Time-Dynamic Profiling with Application to Hospital Readmission Among Patients on Dialysis

Jason Estes, University of Michigan
Danh V. Nguyen, University of California, Irvine
Yanjun Chen, University of California, Irvine
Lorien Dalrymple, University of California, Davis
Connie M. Rhee, University of California, Irvine
Kamyar Kalantar-Zadeh, University of California, Irvine
Damla Senturk*, University of California, Los Angeles

Standard profiling analysis aims to evaluate medical providers, such as hospitals, nursing homes or dialysis facilities, with respect to a patient outcome. Profiling analysis involves regression modeling of a patient outcome, adjusting for patient health status at baseline, and comparing each provider's outcome rate to a normative standard. To date, profiling methods exist only for non time-varying patient outcomes. However, for patients on dialysis, a unique population which requires continuous medical care, methodologies to monitor patient outcomes continuously over time are particularly relevant. Thus, we introduce a novel time-dynamic profiling (TDP) approach to assess the time-varying 30-day readmission rate. TDP is used to estimate, for the first time, the risk-standardized time-dynamic 30-day hospital readmission rate, throughout the time period that patients are on dialysis. We develop the framework for TDP by introducing the standardized dynamic readmission ratio as a function of time and a multilevel varying coefficient model with facility-specific time-varying effects.

EMAIL: dsenturk@ucla.edu

Survival Analysis with Measurement Error in a Cumulative Exposure Variable: Radon Progeny in Relation to Lung Cancer Mortality

Polyna Khudyakov*, Harvard School of Public Health
Jonathan Samet, University of Southern California
Charles Wiggins, University of New Mexico
Xiaomei Liao, Harvard School of Public Health
Angela Meisner, New Mexico Tumor Registry
Donna Spiegelman, Harvard School of Public Health

Exposure variables in occupational and environmental epidemiology are usually measured with error. This error tends to flatten the estimated exposure-response relationship. In this work, we extend the risk set regression calibration (RRC) method for Cox models for cumulative exposure variables to obtain consistent point and interval estimates of relative risks corrected for exposure measurement error. We show that the RRC methodology originally developed for use with an external validation study can be generalized to internal validation study designs as well. We then analyzed the New Mexico uranium miners cohort with follow-up from 1957 to 2012. The exposure data were collected using several different methods of measurement, some of which had a substantial amount of error. After adjusting for bias due to exposure measurement error, the multivariate-adjusted hazard ratio for lung cancer mortality in relation to cumulative radon exposure was estimated to be 4.69[2.21,9.95], substantially

higher than the estimate obtained from the standard analysis ignoring measurement error (1.35[1.21,1.50]). User-friendly software implements this method is publicly available.

EMAIL: stpok@channing.harvard.edu

70. Nonparametric Methods for Functional Data with Application to Clinical Data Analysis

Variable-Domain Functional Regression Models and their Application to the Ecological Momentary Assessment (EMA) Sexually Transmitted Infections (STI) Data

Jaroslaw Harezlak*, Indiana University School of Public Health, Bloomington
Fei He, Institute for Health Metrics and Evaluation
Armando Teixeira-Pinto, University of Sydney

Variable-domain functional regression, a class of scalar-on-function regression models, with the unit-specific observation time is proving to be useful in many observational studies. We extend it to the multiple predictor functions and apply it in the EMA STI study with subject-specific functional predictors. Our data consists of sexual risk behavior and relationship satisfaction measurements obtained in an intensive longitudinal follow-up. We study their relationship with an event defined as a partner change. Specifically, we build an association model for the binary outcome as a time-varying function of functional predictors observed over varying-length time intervals. We extend the existing models from one functional predictor case to multiple predictors. We also extend both cases by including partner-specific random effects quantifying their clustering within the study subjects.

EMAIL: harezlak@iu.edu

Historical Functional Cox Regression, with an Application to Prediction of Multiple Sclerosis Lesions

Elizabeth M. Sweeney*, Rice University
Jonathan Gellar, Mathematica
Ciprian M. Crainiceanu, Johns Hopkins Bloomberg School of Public Health
Philip Reiss, University of Haifa

Historical functional regression was originally developed as a form of function-on-function linear regression: the responses and predictors are defined on a common time domain and the response can be explained by values of the predictor only up to the present time, leading to a bivariate coefficient function supported on a triangular region. Recently, coefficient functions of this type have formed the basis of a novel historical functional Cox model for time-to-event data. The historical Cox model relates the log hazard to not only the present values but also earlier values of a time-varying biomarker. We demonstrate how, by means of this approach, one or more time series of magnetic resonance imaging sequences can serve as predictors of lesion incidence in multiple sclerosis. This application entails sharing information among separate historical models for a set of brain regions, and sheds light on the time lag between early warning signals and clinical manifestation.

EMAIL: ems15@rice.edu

Dynamic Child Growth Prediction: A Comparative Methods Approach

Andrada E. Ivanescu*, Montclair State University
Ciprian M. Crainiceanu, Johns Hopkins Bloomberg School of Public Health
William Checkley, Johns Hopkins University

We introduce a class of dynamic regression models designed to predict the future of growth curves based on their historical dynamics. This class of models incorporates both baseline and time-dependent covariates, start with simple regression models and build up to dynamic function-on-function regressions. We compare the performance of the dynamic prediction models in a variety of signal-to-noise scenarios and provide practical solutions for model selection. We conclude that: 1) prediction

performance increases substantially when using the entire growth history relative to using only the last and first observation; 2) smoothing incorporated using functional regression approaches increases prediction performance; and 3) the interpretation of model parameters is substantially improved using functional regression approaches. Because many growth curve data sets exhibit missing and noisy data we propose a bootstrap of subjects approach to account for the variability associated with the missing data imputation and smoothing. Methods are motivated by and applied to the CONTENT data set, a study that collected monthly child growth data on 197 children from birth until month 15.

EMAIL: ivanescua@mail.montclair.edu

› Variable Selection in the Concurrent Functional Linear Model

Jeff Goldsmith*, Columbia University
Joseph E. Schwartz, Columbia University

We develop methods for variable selection when modeling the association between a functional response and functional predictors that are observed on the same domain. This data structure, and the need for such methods, is exemplified by our motivating example: a study in which blood pressure values are observed throughout the day together with measurements of physical activity, heart rate, location, posture, attitude, and other quantities that may influence blood pressure. We estimate the coefficients of the concurrent functional linear model using variational Bayes and jointly model residual correlation using functional principal components analysis. Latent binary indicators partition coefficient functions into included and excluded sets, incorporating variable selection into the estimation framework. The proposed methods are evaluated in simulated- and real-data analyses.

EMAIL: jeff.goldsmith@columbia.edu

71. Machine Learning Methods

› Evaluating Record Linkage Software Using Representative Synthetic Datasets

Benmei Liu*, National Cancer Institute, National Institutes of Health
Sepideh Mosaferi, University of Maryland

The NCI's Surveillance Epidemiology and End Results (SEER) program has been increasingly engaged in initiatives extensively utilizing record linkage techniques to capture additional medical information (e.g., treatment information, genetic tests, etc.) that cannot be obtained through traditional medical record abstraction. Variety of linkage software products exist and some are free accessible. Challenges arise in choosing the suitable and reliable linkage software and evaluating the linkage quality. Evaluations using real data have restrictions due to unknown truth and limited data accessibility to the patient health identifiers beyond other restrictions related to the divisions and institutes requirements in sharing the dataset. Synthetic but representative datasets may facilitate the evaluation. However, the challenge in simulating data of this type is that errors inserted in the synthetic datasets must be realistic. This paper presents a way to generate synthetic datasets mimicking the US cancer population for record linkage evaluation purpose. Results of evaluating several free accessible record linkage software using the generated synthetic datasets will be shown as well.

EMAIL: liub2@mail.nih.gov

› Semi-Supervised Approaches to Efficient Evaluation of Model Prediction Performance

Jessica Gronsbell*, Harvard University
Tianxi Cai, Harvard University

In many modern machine learning applications, the outcome is expensive or time-consuming to collect while the predictor information is easy to obtain. Semi-supervised learning (SSL) aims at utilizing large amounts of unlabeled data along with small amounts of labeled data to improve the efficiency of a supervised approach. Though numerous SSL classification

procedures have been proposed in recent years, no methods currently exist to evaluate the prediction performance of a working regression model. In the context of developing phenotyping algorithms derived from electronic medical records (EMR), we present an efficient two-step estimation procedure for evaluating a binary classifier based on various prediction performance measures in the semi-supervised (SS) setting. We demonstrate that the proposed estimator is asymptotically normal with variance always smaller than that of its supervised counterpart under correct model specification. Our proposals are also illustrated with an EMR study aiming to develop a phenotyping algorithm for rheumatoid arthritis.

EMAIL: jlg735@mail.harvard.edu

› Group Variable Selection with Compositional Covariates

Anna M. Plantinga*, University of Washington
Michael C. Wu, Fred Hutchinson Cancer Research Center

Feature selection methods for microbiome compositional data are a recent alternative to taxon level or distance-based analyses. Such models effectively handle the high dimensionality of the covariates and can enforce the unit sum constraint of compositional data. However, existing compositional feature selection models do not allow selection at the group level with simultaneous covariate estimation at the taxon level and therefore do not take full advantage of the multi-level structure of microbiome data. We propose an l1/l2 regularized linear log-contrast model that selects groups of covariates and optionally also attains within-group sparsity. This model satisfies the unit-sum constraint for overall and within-group compositions. We express the problem as a constrained convex optimization problem and propose an alternating direction method of multipliers algorithm to fit the model. Selection consistency and bounded loss are guaranteed under fairly weak conditions. The selection and estimation accuracy of our method is evaluated via simulation, and we demonstrate its efficacy by applying it to a study relating host gene expression to gut microbiome composition.

EMAIL: aplantin@uw.edu

› Classification using Ensemble Learning under Weighted Misclassification Loss

Yizhen Xu*, Brown University
Tao Liu, Brown University
Rami Kantor, Brown University
Joseph W. Hogan, Brown University

Binary classification rules based on covariates typically depend on simple loss functions such as zero-one misclassification. Some cases may require more complex loss functions. HIV monitoring of people on antiviral treatment requires periodic assessment of treatment failure, viral load (VL) above a certain level. In resource limited settings, VL tests may be limited by cost or technology, and diagnoses are based on other clinical markers. Higher premium is placed on avoiding false-positives which brings greater cost and reduced treatment options. Here, the optimal rule is determined by minimizing a weighted misclassification loss. We propose a method for finding and cross-validating optimal binary classification rules under weighted misclassification loss. We focus on rules comprising a prediction score and an associated threshold, where the score is derived using ensemble learner. Simulations and applications suggest that our method, which derives the score and threshold jointly, more accurately estimates overall risk and has better operating characteristics compared to two-step methods that derive the score first and the cutoff conditionally on the score.

EMAIL: yizhen_xu@brown.edu

› FDR Control of the High Dimensional TOST Tests

Yue Qi, University of Delaware
Jing Qiu*, University of Delaware

High dimensional equivalence testing is a very important but seldom studied problem. When researchers look for equivalently expressed genes, the common practice is to conduct dierential tests and treat genes that are not dierentially expressed as equivalently expressed genes. This is statistically not valid because it does not control the type I error appropriately. An appropriate way is to conduct equivalence tests. A well-known equivalence test is two one-sided tests (TOST). The existing FDR controlling methods are overconservative for

equivalence tests. We investigate the performance of exisitng FDR controlling methods and propose three new methods to control the FDR for equivalence test.

EMAIL: Qiujing@udel.edu

› Evolutionary State Space Model and its Application to Time-Frequency Local Field Potentials Analysis

Xu Gao*#, University of California, Irvine
Hernando Ombao, University of California, Irvine

We develop a model for high dimensional signals driven by sources whose properties evolve over epochs in the experiment. We propose the evolutionary state space model(E-SSM)where signals are mixtures of sources having oscillatory activity at defined frequency bands. One unique feature in E-SSM is that the sources are parametrized as second order autoregressive AR(2) processes. To account for non-stationarity of sources, the AR(2) parameters are allowed to vary over epochs. In contrast to independent component analysis, our method accounts for the temporal structure of the sources. Compared to the data-adaptive strategy such as filtering, the proposed E-SSM easily accommodates non-stationarity through source parameters. To estimate the model, we use Kalman smoother, maximum likelihood and blocked resampling approaches. The E-SSM model is applied to a multi-epoch LFP signals from a rat performing a non-spatial(olfactory) encoding task. Our method captures the evolution of the power of different sources across encoding phases of the experiment. The E-SSM model also identifies clusters of tetrodes that behave similarly with respect to the decomposition of the different sources.

EMAIL: xgao2@uci.edu

72. Varial Subset Selection/Model Selection

› Decomposition-Gradient-Regression Approach for High-Dimensional Correlated Data with Binary Outcome

Yuanzhang Li, Walter Reed Army Institute of Research
Hailin Huang*, The George Washington University
Hua Liang, The George Washington University
Colin Wu, National Heart, Lung and Blood Institute, National Institutes of Health

In this research, we examine the Decomposition-Gradient-Regression (DGR) method to identify significant biomarkers when the collected biomarkers are strongly correlated and to make predictions. The method mainly consists of three steps: (i) Decompose the space of biomarkers into subspaces such that all biomarkers in each subspace are independent; (ii) In each subspace, find a gradient that is a linear combination of biomarkers, such that the gradient optimally separate the binary responses; (iii) Use gradient regression method to identify significant biomarkers in each subspace and to make predictions. We investigate the properties of DGR and compare it with LASSO type methods through simulation studies. Finally, the DGR method is applied to determine potential association of 64 biomarkers from a bipolar study.

EMAIL: hhl1988@gwu.edu

› Sequential Estimation in Sparse Factor Regression

Aditya Kumar Mishra*, University of Connecticut
Dipak Dey, University of Connecticut
Kun Chen, University of Connecticut

Multivariate regression models are increasingly required and formulated in various fields. A sparse singular value decomposition of the regression component matrix is appealing for reducing dimensionality and facilitating interpretation, but its recovery remains a challenging problem computationally. We formulate the problem as a sparse factor regression and develop an efficient sequential computation procedure. At each sequential step, a latent factor is constructed as a sparse linear combination of the observed predictors, for predicting the responses after accounting for the effects of

the previous factors; each factor is allowed to potentially influence only a subset of responses. Each sequential step reduces to a regularized unit-rank regression in which the orthogonality constraints among the sparse factors become optional rather than necessary. Coordinate descent and Lagrangian multipliers are utilized to ensure fast computation and algorithmic convergence, even in the presence of missing data. We justify our approach by showing oracle properties of the estimator. The efficacy of our method is demonstrated by simulation studies and two real applications in genetics.

EMAIL: aditya.mishra@uconn.edu

› Bayesian Indicator Variable Selection Model with Multi-Layer Overlapping Groups

Li Zhu*, 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. In this paper, 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 (such as its mRNA expression, copy number variation and methylation levels). In addition to the multi-layer structure, the groups may be overlapped. We propose a Bayesian indicator model that can conveniently incorporate the multi-layer overlapping group structure in variable selection. We discuss properties of the proposed prior and prove selection consistency and asymptotic normality of the posterior median estimator of the method. We apply the model to three simulations and one TCGA breast cancer example to demonstrate its superiority over other existing methods. The result not only enhances prediction accuracy but also improves variable selection and model interpretation.

EMAIL: liz86@pitt.edu

› Nonconvex Penalized Regression using Depth-Based Penalty Functions: Multitask Learning and Support Union Recovery in High Dimensions

Subhabrata Majumdar*, University of Minnesota
Snigdhansu Chatterjee, University of Minnesota

We propose a new class of nonconvex penalty functions in the paradigm of penalized sparse regression that are based on data depth-based inverse ranking. Focusing on a one-step sparse estimator of the coefficient matrix that is based on local linear approximation of the penalty function, we derive its theoretical properties and provide the algorithm for its computation. For orthogonal design and independent responses, the resulting thresholding rule enjoys near-minimax optimal risk performance, similar to the adaptive lasso (Zou, 2006). A simulation study as well as real data analysis demonstrate its effectiveness compared to present methods that provide sparse solutions in multivariate regression.

EMAIL: zoom.subha@gmail.com

› Variable Selection from Clusters of Predictors via Model Free Surrogate Variable Construction

Haileab Hilafu*, University of Tennessee
Wenbo Wu, University of Oregon

Many data sets consist of predictors with an inherent grouping structure. We are interested to select at least one variable from each group in the context of predictive regression modeling. To this end, we present a two-stage variable selection method. In the first stage we use model free sufficient dimension reduction methods to construct surrogate variable for each group. This surrogate variable is model free and accounts for potential non-linear dependence between the response and the predictors. This first step also yields aggregate weight vector that measures the overall contribution of variable towards predicting the response. In the second step, with the aid of this aggregate weight vector, we fit an adaptive lasso of the surrogate variable in the predictors to obtain a sparse estimate of the weight vector for the overall contribution of the predictors. Extensive simulation study shows that the propose method is more robust to potential non-linear, and applicable to multi-index regression models.

EMAIL: hhilafu@utk.edu

Variable Selection for Generalized Single Index Coefficient Model

Shunjie Guan*, Michigan State University
Xu Liu, Shanghai University of Finance and Economics
Yuehua Cui, Michigan State University

Scientific research suggests the joint effect of environmental and genetic factors (G $\times$ E) on disease risk, however, its effect on complicated disease remain largely unknown. For this purpose, we proposed to use generalized single index coefficient model (GSICM) to identify how several environmental factors as a whole modulate gene expressions. GSICM is a semiparametric model that allows nonlinear G $\times$ E interaction. We proposed a three steps penalized variable selection approach for GSICM that not only classify the non-parametric functions into three categories: varying, constant or zero, but also select the non-zero loading parameters. The non-parametric functions are approximated by B-spline basis function, and its coefficients are estimated using group coordinate descent algorithm, while the loading parameters are estimated using local quadratic approximation algorithm. Under some regularity conditions, the proposed method is shown to possess oracle property. The performance of the proposed variable selection approach are evaluated in both selection and estimation accuracy via simulation studies and the model is further implemented to a real data example.

EMAIL: guanshun@stt.msu.edu

73. Spatial/Temporal Modeling

A New Class of Statistical Models for Longitudinal Network Data

Ming Cao*, University of Texas Health Science Center at Houston
Yong Chen, University of Pennsylvania
Kayo Fujimoto, University of Texas Health Science Center at Houston
Michael Schweinberger, Rice University

Longitudinal network data provide a rich source of data for studying link formation processes over time. Such data often exhibit both community structure - which may evolve over time and may be unobserved - and complex past-present and present-present network dependencies (e.g., transitivity). To capture unobserved, time-evolving community structure as well as complex network dependencies, we propose a novel class of statistical models called Temporal Hierarchical Exponential Random Graph Models (THERGMs). To estimate THERGMs, we follow a two-step estimation approach. In the first stage, we estimate blocks. In the second stage, we estimate TERGMs within and between blocks. We present simulation results and applications.

EMAIL: com.gdcc.hp.caoming@gmail.com

Treatment Dependent Covariance: Modeling and Inference

Emily Leary*, University of Missouri
Alison Gerken, Agricultural Research Service, U.S. Department of Agriculture
Lauren M. McIntyre, University of Florida
Alison Morse, University of Florida
Patrick J. Brown, University of Illinois, Urbana-Champaign
Andrew D. B. Leakey, University of Illinois, Urbana-Champaign
Elizabeth A. Ainsworth, University of Illinois, Urbana-Champaign
Linda J. Young, National Agricultural Statistics Service, U.S. Department of Agriculture

The problem of differential covariance among treatments may exist in many scenarios but is often ignored. We discuss the statistical challenges of treatment-induced covariance in the context of a project whose goal was to provide an understanding of genetic variation in maize response to atmospheric ozone concentration. In this experiment, complexity in assessing the effect of ozone is due to spatial heterogeneity in growing conditions across the experimental field, wind effects that produce differential application of the ozone treatment within a replicate plot, and a small number of replicates. Re-randomization of the field location for genotypes was completed to produce replicate measurements, but these were located in different places and genotypes could not be planted in all positions relative to the point-of-release of ozone. Discussion will focus on the effects of incorporating a treatment-dependent covariance structure. We find that more conservative inferences are made when accounting for heterogeneity of exposure and posit that such treatment dependent

covariances may be more widespread. We show a straightforward way to model these effects.

EMAIL: learye@health.missouri.edu

› A Blockmodel for Node Popularity in Networks with Community Structure

Srijan Sengupta*, Virginia Tech
Yuguo Chen, University of Illinois, Urbana-Champaign

Network data analysis is a fast growing research field with diverse applications spanning several scientific disciplines. The community structure observed in empirical networks has been of particular interest in the statistics literature, with a strong emphasis on the study of blockmodels. In this paper we study an important network feature called node popularity, which is closely associated with community structure. Neither the classical stochastic blockmodel nor its degree-corrected extension can satisfactorily capture the dynamics of node popularity as observed in empirical networks. We propose a popularity-adjusted blockmodel for flexible and realistic modeling of node popularity. We establish consistency of likelihood modularity for community detection, as well as estimation of node popularities and model parameters, and demonstrate the advantages of the new modularity over the degree-corrected blockmodel modularity in simulations. By analyzing the political blogs network, the British MP network, and the DBLP bibliographical network, we illustrate that improved empirical insights can be gained through this methodology.

EMAIL: sengupta@vt.edu

› A Novel Framework for Spatially-Varying Age-Period-Cohort Analysis with Application to U.S. Mortality, 1999-2014

Pavel Chernyavskiy*, National Cancer Institute, National Institutes of Health
Mark P. Little, National Cancer Institute, National Institutes of Health
Philip S. Rosenberg, National Cancer Institute, National Institutes of Health

Age-period-cohort models decompose trends in population rates into age, calendar-period, and birth-cohort effects. Methodology to fit parsimonious models to data in geographically-organized regions (e.g., US states) is not well developed. Here, we present a general method for modeling trends in geographically-organized regions. We allow region-specific parameters to be correlated spatially (e.g., among neighboring states), and to one another (e.g., between baseline risk and calendar-period trend), via a random-effects formulation using a generalized multivariate conditionally auto-regressive prior implemented using a Gibbs sampler in JAGS. We apply our approach to US state-level mortality in young (aged 25-50) white non-Hispanic men and women to assess the impact of fatal drug overdoses (OD) on total mortality. We show that OD fully accounts for rising mortality among women; and that increasing mortality is positively correlated with baseline risk in men and women, suggesting that the disparities between states have grown over time. Our model parsimoniously accounts for spatial heterogeneity in model parameters, providing reliable inference in large national databases.

EMAIL: pavel.chernyavskiy@nih.gov

› Parametric Accelerated Failure Time Model with Spatially Varying Coefficients

Guanyu Hu*, Florida State University
Fred Huffer, Florida State University
Debajyoti Sinha, Florida State University
Jonathan Bradley, Florida State University

Accelerated failure time model is an universally used model in the survival analysis. In the public health study, people often collect the data from different locations of the medical services provider. There are large geographical variations in survival rate of cancer data. In this paper, we foucused on the accelerated failure time model with spatially varying coefficients. We proposed approximated estimation method of our model by using MCMC. In addition, we used krigin method to predict the surivival model of location without observation. Finally, we applied the our model into prostate cancer data of Louisiana from the SEER program to show the spatially varying effects on survival rates of prostate cancer.

EMAIL: guanyu.hu@stat.fsu.edu

› Modelling Multivariate Spatial Processes Using Data from Different Support

Kelly-Ann Dixon Hamil*, Purdue University
Hao Zhang, Purdue University

One of the main aims of statistical analysis is to make meaningful predictions, the accuracy of which may be achieved by including related variables in the model. The implementation of this may become cumbersome when related data are measured at different levels of support and/or are collected from different sources. Using data for biomass, collected at the area level, and temperature, collected at the point level, we will suggest a technique using low rank methodology and semiparametric covariance functions to model the covariance function which will in turn produce improved predictions for biomass.

EMAIL: dixon13@purdue.edu

74. Biomarkers

› Joint Analysis of Left-Censored Longitudinal Biomarker and Binary Outcome via Latent Class Modeling

Menghan Li*, The Pennsylvania State University
Lan Kong, The Pennsylvania State University

Joint latent class modeling is an appealing approach for joint analysis of longitudinal biomarkers and clinical outcomes when the study population is heterogeneous. The link between the marker trajectory and the risk of event is reflected by the latent classes, which accommodate the underlying population heterogeneity. Since joint latent class models are essentially mixture models, parameter estimation is challenging as the number of parameters increases. We develop a new Monte Carlo EM (MCEM) algorithm based on Metropolis-Hasting method for joint analysis of a biomarker and a binary outcome in latent class modeling framework. To account for left-censoring in the biomarker measurements due to a lower detection limit, we modify the joint likelihood and demonstrate the utility of MCEM to incorporate the censored biomarker data. We conduct simulation studies to evaluate the performance of our MCEM algorithm and apply the proposed method to a sepsis study to determine if the pattern of cytokine trajectory is associated with the 90-day mortality, and whether there exists subpopulations with differential cytokine profiles and mortality risks.

EMAIL: mul283@psu.edu

› Bayesian Wavelet Shrinkage with Beta Priors for Curve Estimation with Lognormal Positive Errors

Nancy L. Garcia*, University of Campinas
Alex R. Sousa, University of Campinas

Changes in the lipid composition of the bovine uterus exposed to greater (LF-LCL group; n = 7) or lower (SF-SCL group/ n = 10) concentrations of progesterone during post-ovulation were investigated by matrix assisted laser desorption ionization-mass spectrometry (MALDI-MS) and raw spectra data were obtained. Wavelet shrinkage methods have been extremely useful to reduce noise present in the data and to estimate curves by expansion in wavelet basis. Typically, these methods shrink to zero empirical coefficients sufficiently close to zero. In this paper, we propose a Bayesian wavelet shrinkage rule that assumes a shifted beta prior distribution to the true wavelet coefficients for noise reduction and curve estimation assuming positive distribution for the error. We provide the performance of the rule in simulated data sets based on the Donoho and Johnstone test functions and comparisons with shrinkage rules already used.

EMAIL: nancy@ime.unicamp.br

› A Mixture of Bivariate Truncated Beta Distribution Applied to the Classification of Matrix Assisted Laser Desorption Ionization Mass Spectrometry (MALDI-MS) Data

Mariana R. Motta*, University of Campinas
Nancy L. Garcia, University of Campinas

Changes in the lipid composition of the bovine uterus exposed to greater (LF-LCL group; n = 7) or lower (SF-SCL group/ n = 10) concentrations of progesterone during post-ovulation were investigated by matrix assisted laser desorption ionization-mass spectrometry (MALDI-MS). For each cow two measurements were made. After preprocessing

of the data, 77 {\it m/z} values were selected identifying specific ions in the spectra. Due to the small sample size, the usual methods could not identify a biomarker that discriminated between groups. A model-based approach was therefore proposed and found able to classify the MALDI-MS data, fitting a mixture of bivariate beta distributions truncated to accommodate the large number of zero observations as well as the bivariate nature of the data.

EMAIL: marianar@ime.unicamp.br

› Linear Combinations of Diagnostic Markers for a Specific Clinical Application

Andriy Bandos*, University of Pittsburgh
David Gur, University of Pittsburgh

Linear combination of several markers is often used to produce an explicit decision rule for improved diagnosis of medical conditions. These combinations are often optimized by maximizing an objective function related to diagnostic accuracy (e.g., the area under the ROC curve, AUC). However, global indices, such as the AUC, are often too general for applications constrained by clinical practice considerations. A common approach to address this issue is to optimize the partial AUC (pAUC) over the clinically relevant range (e.g. high specificity), with the underlying assumption that the ultimately resulting decision rule would have optimal sensitivity in the targeted range. We investigated the maximum sensitivity that can be achieved in a high specificity range (Se|sp0) when using different optimization targets. We demonstrated that optimizing for the pAUC in many scenarios leads to Se|sp0 that is substantially worse than under the AUC-based optimization. Both optimizations can be substantially outperformed by directly maximizing Se|sp0. Thus, when AUC is deemed too crude, instead of optimizing toward pAUC the optimization of marker combination should be focused directly on Se|sp0.

EMAIL: anb61@pitt.edu

› Optimal Decision Rule for Multiple Biomarkers Combined as Tree-Based Classifiers

Yuxin Zhu*, Johns Hopkins University
Mei-Cheng Wang, Johns Hopkins University

In biomedical practices, multiple biomarkers are often combined using a classification rule with tree structure to make diagnostic decisions. The classification structure and cutoff point at each node of a tree are commonly chosen based on experience of decision makers. There is a lacking of analytical approaches that lead to optimal prediction performance, or to guide the choice of optimal cutoff points of a pre-specified classification tree. In this paper we propose to search for and estimate the optimal decision rule through an approach of rank correlation maximization. The proposed method is flexible and computationally feasible when there are many biomarkers available for classification or prediction. Using this method, for a pre-specified classification rule, we are able to guide the choice of optimal cutoff at tree nodes, as well as estimate optimal prediction performance of multiple biomarkers combined by a tree-based classification rule.

EMAIL: daisy.zhu.yx@gmail.com

› Asymptotic Distribution of Delta AUC, NRI, and IDI Based on U-Statistics Theory

Olga V. Demler*, Brigham and Women's Hospital
Michael Pencina, Duke University
Nancy R. Cook, Brigham and Women's Hospital
Ralph B. D'Agostino, Boston University

The change in AUC (delta AUC), the IDI and NRI are commonly used measures of predictive model performance. However, a risk prediction model enhanced with uninformative predictor(s) can lead to a significant improvement of NRI, and a model enhanced with a strong predictor can fail to produce a significant increase in AUC. Several explanations of these phenomena have been suggested, i.e. incorrect variance estimator, lack of training-validation approach and a flaw in the statistics itself. In this paper we unite the delta AUC, IDI and NRI under the umbrella of U-statistics family. Using powerful statistical theory developed for U-statistics, we explained several reported contradictions in their behavior. We proved that delta AUC, IDI and three types of NRI asymptotically follow normal distribution, unless they compare nested models under the null. In the latter case delta AUC, NRI and IDI are non-normal and most CI formulas are invalid. Using results of Randles, de Wet and Sukhatme we discuss, when their vari-

ance formulas should be adjusted for estimated parameters and when such adjustment is unnecessary.

EMAIL: olgademler@gmail.com

75. Presidential Invited Address

› But I'm a Data Scientist Too, Aren't I?

Louise Ryan, Ph.D., Distinguished Professor, School of Mathematical and Physical Sciences, University of Technology Sydney

The statistics profession is in a period of disruptive change, heralded by explosive growth in information technology and the "big data" revolution. New specialties such as machine learning, data science and analytics seem to go from strength to strength and sometimes it seems like statistics is being discarded like one of last decade's fashion embarrassments. In this presentation, I will offer some perspectives on the changing landscape for statistical science. I'll draw on some of my own recent projects where "statistics as usual" falls short and outline some of the areas where I think there are great opportunities for our profession to strengthen our role in the data science arena. I'll finish up with some thoughts about implications for training the next generation as well as up skilling the current generation of statisticians.

EMAIL: Louise.M.Ryan@uts.edu.au

76. A Life Too Short: Dan Sargent as Leader, Researcher, Colleague and Mentor

› Tribute to Daniel Sargent: A Friend and Clinical Researcher's Perspective

Richard M. Goldberg*, The Ohio State University School of Medicine

Dan Sargent was a consummate collaborator. He worked with researchers, government agencies, industry, professional organizations, and academic statisticians worldwide. As a voice of reason, innovation and collegiality he was widely sought after and recruited into leadership roles. He was the lead statistician of multiple cancer clinical trials groups including the North Central Cancer Treatment Group, American College of Surgeons Oncology Group and the Alliance for Clinical Trials in Oncology. He chaired Subcommittee H, the NCI group charged with evaluating NCI designated Clinical Trials Organizations, as well as serving on many other committees and working groups. He was an advisor and DMC member for many companies, assisting them in drug development. He was elected president of the Society for Clinical Trials and worked on many activities for the American Society of Clinical Oncology and other professional organizations. We had the opportunity to write 90 papers together with collaborators from around the world over the 20 years of our professional relationship and friendship, and helped to nurture the careers of many clinicians and statisticians.

EMAIL: Richard.goldberg@osumc.edu

› Dan Sargent as Leader, Researcher, Colleague and Mentor

Marc Buyse*, International Drug Development Institute (IDDI)

I will discuss Dan Sargent's contributions to the validation of surrogate endpoints in oncology, both on the theoretical and applied fronts. One of Dan's most frequently cited papers was his analysis of a large meta-analysis of clinical trials in patients with resectable colon cancer [J Clin Oncol 2005; 23:8664-70]. In this paper, he showed that 3-year disease-free survival is a valid surrogate for 5-year overall survival in the evaluation of new treatments for these patients, thus allowing new treatments to be approved 2 years earlier if disease-free survival is used as a primary endpoint, instead of overall survival. This landmark paper attracted considerable attention at the Food and Drug Administration, and started a new era of search for surrogates in other tumor types. On the methodological front, Dan and his colleagues contributed to the development of new methods for surrogate endpoint evaluation [CSDA 2011; 55:2748-57], using such diverse methods as copulas, Bayesian models, accelerated failure time models and causal inference. Dan left an extraordinarily rich and informative legacy in this emerging area of research.

EMAIL: marc.buyse@iddi.com

Dan Sargent's Contributions to the National Cancer Institute

Edward L. Korn*, National Cancer Institute, National Institutes of Health

Dan Sargent was a tireless champion for having appropriate statistical methods used for evaluating cancer treatments. This can be seen in his many contributions to the National Cancer Institute. These include his work as an NCI Cooperative Group statistician (for the North Central Cancer Treatment Group, American College of Surgeons Oncology Group and the Alliance for Clinical Trials in Oncology), and as a member of many NCI task forces (multiple task forces for the NCI Colon Cancer Steering Committee and the Clinical Trial Design Task Force of the NCI Investigational Drug Steering Committee). He was also a member of various NCI review committees (including chairing Subcommittee H and being a member of the NCI Biometrics Methods study section) and NCI advisory committees (including the Clinical Trials Advisory Committee). His selfless giving of his time for these endeavors has improved the lives of cancer patients.

EMAIL: korne@ctep.nci.nih.gov

Dan Sargent: A Wonderful Mentor and a Great Man

Jared C. Foster*, Mayo Clinic

A little over a year ago, I joined the faculty in the Section of Cancer Center Statistics at Mayo Clinic, and since that time, Dan Sargent has been my primary mentor. Over the past year, I have been completely transformed, both as a statistician and as a person. Dan was an extremely talented statistician and statistical mentor, and through his words and his actions, he taught me how to thrive in a fast-paced, collaborative environment, and helped me find myself as a statistician. More importantly, Dan was a perfect example of what I hope to become personally. He was incredibly kind and thoughtful, and just being around him was enough to brighten my day. He loved his family very much, and seemed to always put them first. He was the greatest mentor I've ever had, and I will miss him terribly. In this talk, I will give my perspective on what it was like to work for and learn from Dan, and in addition, I will share thoughts and stories from a number of my colleagues at Mayo Clinic.

EMAIL: foster.jared@mayo.edu

77. Heterogeneous Modeling and Individual Learning in Longitudinal Data

High-Dimensional A-Learning for Optimal Dynamic Treatment Regimes

Chengchun Shi, North Carolina State University
Rui Song, North Carolina State University
Wenbin Lu*, North Carolina State University

In this work, we propose a penalized multi-stage A-learning for deriving the optimal dynamic treatment regime when the number of covariates is of the non-polynomial (NP) order of the sample size. To preserve the double robustness property of the A-learning method, we adopt the Dantzig selector which directly penalizes the A-leaning estimating equations. Oracle inequalities of the proposed estimators for the parameters in the optimal dynamic treatment regime and error bounds on the difference between the value functions of the estimated optimal dynamic treatment regime and the true optimal dynamic treatment regime are established. Empirical performance of the proposed approach is evaluated by simulations and illustrated with an application to data from the STAR*D study.

EMAIL: lu@stat.ncsu.edu

GLIMMIX with Latent Gaussian Mixture: Theory and Applications

Yi Li*, University of Michigan

Numerical analyses reveal that the distributional assumptions in GLIMMIX do matter especially when the effect is extreme. We propose to model the latent random variable with a finite Gaussian mixture distribution. Such latent Gaussian mixture models provide a convenient framework to study the heterogeneity among different clusters. To overcome the weak identifiability issues, we propose to estimate the latent Gauss-

ian mixture model using a penalized likelihood approach, and develop sequential locally restricted likelihood ratio tests to determine the number of components in the Gaussian mixture distribution.

EMAIL: yili@umich.edu

› Change-Point and Chromosome Copy Number Detection in Multiple Samples

Chi Song, The Ohio State University
Xiaoyi Min, Georgia State University
Heping Zhang*, Yale University

Chromosome copy number variation (CNV) is the deviation of genomic regions from their normal copy numbers that may associate with many human diseases. CNVs can be called by detecting the change-points in mean for sequences of array-based intensity measurements. The majority of the available CNV calling methods are single sample based. Only a few multiple sample methods have been proposed using scan statistics that are computationally intensive and designed toward either common or rare change-points detection. We propose a novel multiple sample method by adaptively combining the scan statistic of the screening and ranking algorithm, which is computationally efficient and is able to detect both common and rare change-points. We show that asymptotically this method can find the true change-points with almost certainty and multiple sample methods are superior to single sample methods when shared change-points are of interest. We demonstrate the superority of our proposed method to competing approaches by detecting CNVs from the Primary Open-Angle Glaucoma Genes and Environment study.

EMAIL: heping.zhang@yale.edu

› Exploration of Longitudinal Features via Semiparametric Mixture Models

Naisyin Wang*, University of Michigan

In this presentation, we couple the uses of semiparametric approaches and latent mixture variables to explore the underlying features of longitudinal data. Researchers in various fields have used mixture models to accommodate subsets heterogeneity and unequal levels of noises. Non- and semiparametric modeling is known to be useful in structure exploration. We utilize the strengths of these two modeling strategies to reflect the data complexity both locally and globally. We will focus on exploration as well as on prediction. We aim at extracting different sub-features in the data, to best reflect the shared and contrast information embedded in different sub-groups of subjects. Criteria, including out-of-sample prediction, were employed to gauge the use of different types of features and the bases on which they were evaluated. Effectiveness of the new methods is demonstrated using synthetic data and data collected through medical studies.

EMAIL: nwangstat@gmail.com

78. Advanced Bayesian Nonparametric Methods for Biomedical Studies

› High-Dimensional Spatial-Temporal Models for Mapping Environmental Pollutants

Sudipto Banerjee*, University of California, Los Angeles
Abhirup Datta, Johns Hopkins University
Lu Zhang, University of California, Los Angeles
Andrew O. Finley, Michigan State University

With the growing capabilities of GIS, statisticians today routinely encounter massive amounts of space-time data. Bayesian hierarchical spatial-temporal process models are widely deployed to better understand the complex nature of spatial and temporal variability. However, fitting these models involves expensive matrix computations with cubic order complexity, rendering them impracticable. I will particularly focus upon sparsity-inducing Nearest-Neighbor Gaussian Process (NNGP) that can be embedded within a rich and flexible hierarchical modeling framework and some computational strategies using Hamiltonian Monte Carlo to deliver exact Bayesian inference. The talk will demonstrate its use in inferring on the spatial-temporal distribution of ambient air pollution in continental Europe using spatial-temporal regression models with the LOTUS chemistry transport models.

EMAIL: sudipto@ucla.edu

› Scalable Nonparametric Bayes for Complex Biomedical Data

David Dunson*, Duke University

We propose new classes of models and corresponding scalable computational algorithms that can be used broadly in biomedical applications.

EMAIL: dunson@stat.duke.edu

› Bayesian Nonparametric Hypothesis Tests

Chris C. Holmes*, University of Oxford
Sarah Filippi, University of Oxford

Hypothesis testing, such as two-sample tests and tests for independence, are important components of biomedical data analysis. We discuss recent advances in the use of Bayesian nonparametric (BNP) models for this task. One advatange of BNP methods is that they provides an explicit model based probability measure for the null hypothesis Pr(H0 | data), as well as for the alternative, Pr(H1 | data) = 1 - Pr(H0 | data), without making parametric assumptions on the form of the likelihood function. This allows one to encorporate prior information and control for multiple testing, while naturally extending to tests for contrasts where we wish to detect for an effect under one condition but not under another. This latter task is problematic for non-Bayesian methods as one of the tests has a higher dimension null than the alternative, but is readily accomodated in the Bayesian approach. We present some examples from biomedical genomics.

EMAIL: cholmes@stats.ox.ac.uk

› Bayesian Inference for Latent Biologic Structure with Determinantal Point Processes (DPP)

Peter Mueller*, University of Texas, Austin
Yanxun Xu, Johns Hopkins University
Donatello Telesca, University of California, Los Angeles

We discuss the use of the determinantal point process (DPP) as a prior for latent structure in biomedical applications, where inference often centers on the interpretation of latent features as biologically or clinically meaningful structure. Typical examples include mixture models, when the terms of the mixture are meant to represent clinically meaningful subpopulations (of patients, gene's, etc.). Another class of examples are feature allocation models. We propose the DPP prior as a repulsive prior on latent mixture components in the first example, and as prior on feature-specific parameters in the second case. We argue that the DPP is in general an attractive prior model for latent structure when biologically relevant interpretation of such structure is desired. An important part of our argument are efficient and straightforward posterior simulation methods. We implement a variation of reversible jump Markov chain Monte Carlo simulation for inference under the DPP prior, using a density with respect to the unit rate Poisson process.

EMAIL: pmueller@math.utexas.edu

79. Modeling Complexities in Spatial and Spatio-Temporal Data

› Modeling Bronchiolitis Epidemics in the Presence of Spatio-Temporal Uncertainty

Matthew J. Heaton*, Brigham Young University
Sierra Pugh, Brigham Young University
Candace Berrett, Brigham Young University
Brian Hartman, Brigham Young University
Chantel Sloan, Brigham Young University

Bronchiolitis is the most common cause of infant hospitalization in the United States and results in approximately 500 infant deaths annually. Because no vaccine currently exists for bronchiolitis, prevention strategies hinge on epidemiological research of bronchiolitis and its methods of spread, reasons for seasonal behavior, and response to different environmental, climatological and meteorological variables. To better understand bronchiolitis, the US Military Health System (MHS) compiled a bronchiolitis database of more than 140,000 cases that have been spatially and temporally randomized (jittered) to ensure privacy. Although such a large amount of data has the potential to further epidemiological understanding of bronchiolitis, if not properly accounted for, the spatial uncertainty due to the jittered observations can bias analyses. This research develops a statistical change point model for the bronchiolitis cases that estimates the start and peak times of bronchiolitis epidemics across the contiguous US.

Importantly, the randomization of the spatial and temporal locations are appropriately built into the model structure.

EMAIL: mheaton@stat.byu.edu

› Dynamic Multiscale Spatiotemporal Models for Poisson Data

Thais C. O. Fonseca, Federal University of Rio de Janeiro
Marco A. R. Ferreira*, Virginia Tech

We propose a new class of dynamic multiscale models for Poisson spatiotemporal processes. Specifically, we use a multiscale spatial Poisson factorization to decompose the Poisson process at each time point into spatiotemporal multiscale coefficients. We then connect these spatiotemporal multiscale coefficients through time with a novel Dirichlet evolution. Further, we propose a simulation-based full Bayesian posterior analysis. In particular, we develop filtering equations for updating of information forward in time and smoothing equations for integration of information backward in time, and use these equations to develop a forward filter backward sampler for the spatiotemporal multiscale coefficients. Because the multiscale coefficients are conditionally independent a posteriori, our full Bayesian posterior analysis is scalable, computationally efficient, and highly parallelizable. Finally, we illustrate the usefulness of our multiscale spatiotemporal Poisson methodology with two applications. The first application examines mortality ratios in the state of Missouri, and the second application considers tornado reports in the American Midwest.

EMAIL: marf@vt.edu

› Understanding the Mortality Burden of Air Pollution in Michigan Accounting for Preferential Sampling

Veronica J. Berrocal*, University of Michigan
Carina Gronlund, University of Michigan
David M. Holland, U.S. Environmental Protection Agency

The deleterious effect of exposure to ambient air pollution on human health has been recognized in several environmental epidemiological studies. Linking ambient air pollution exposure to health outcomes is often problematic as ambient air pollution is not measured everywhere or where health data is available. To address this issue, several strategies have been proposed, including using spatial statistical models to predict (or impute) the unobserved air pollution levels at the needed spatial locations and resolution. However, most of these model do not take into account the fact that air quality monitors are preferentially located at sites where pollution levels are expected to be high (preferential sampling). In this talk, we focus on fine particulate matter, as air pollutant, and mortality in Michigan during years 2001-2010, as health outcome, and we propose spatio-temporal statistical models that: (i) address the preferential placement of monitors, (ii) combine different data sources on air pollution, and (iii) evaluate the impact of the preferential placement of air quality monitors on the estimated mortality burden of PM2.5 in Michigan over a 10-year period.

EMAIL: berrocal@umich.edu

› Geostatistical Methods for Massive Imaging Data

Brian Reich, North Carolina State University
Joseph Guinness*, North Carolina State University
Taki Shinohara, University of Pennsylvania
Ana Maria Staicu, North Carolina State University
Simon Vandekar, University of Pennsylvania

Imaging data arises in diverse scientific fields, such as neuroscience and remote sensing of atmospheric and oceanic processes. Modern automatic monitoring instruments have the ability to collect massive quantities of data and can transform our understanding of physical and biological processes. It is a challenge to analyze these massive datasets without making overly simplistic model assumptions. In this talk, we discuss modeling of massive and spatially correlated datasets and computational approaches for analyzing them based on spectral representations of the models. The approaches are applied to brain imaging data measuring cortical thickness.

EMAIL: jsguinne@ncsu.edu

80. Quantile Regression and Inference for High Dimensional Data Analysis

Principal Quantile Regression for Sufficient Dimension Reduction with Heteroscedasticity

Chong Wang, North Carolina State University
Seung Jun Shin, Korea University
Yichao Wu*, North Carolina State University

Sufficient dimension reduction (SDR) is a successive tool for reducing data dimensionality without stringent model assumptions. In practice, data often display heteroscedasticity which is of scientific importance in general but frequently overlooked since a primal goal of most existing statistical methods is to identify conditional mean relationship among variables. In this article, we propose a new SDR method called principal quantile regression (PQR) that efficiently tackles heteroscedasticity. PQR can naturally be extended to a nonlinear version via kernel trick. Asymptotic properties are established and an efficient solution path-based algorithm is provided. Numerical examples based on both simulated and real data demonstrate the PQR's advantageous performance over existing SDR methods. PQR still performs very competitively even for the case without heteroscedasticity.

EMAIL: wu@stat.ncsu.edu

Variable Selection for High-Dimensional Additive Quantile Regression

Ben Sherwood*, University of Kansas
Lan Wang, University of Minnesota
Adam Maidman, University of Minnesota

Additive quantile regression is used to estimate a conditional quantile with few assumptions about the relationship between the response and the predictors. To estimate such a model in the high-dimensional setting we propose using B-splines to estimate the potentially nonlinear relationships and a non-convex group penalty for variable selection. Estimation and variable selection properties of the method will be presented. In addition, an algorithm for approximating the nonsmooth and nonconvex objective function is presented.

EMAIL: ben.sherwood@ku.edu

High Dimensional Censored Quantile Regression

Qi Zheng*, University of Louisville
Limin Peng, Emory University
Xuming He, University of Michigan

Censored quantile regression (CQR) has emerged as a useful regression tool for survival analysis. Some commonly used CQR methods can be characterized by stochastic integral-based estimating equations in a sequential manner across quantile levels. In this work, we analyze CQR in a high dimensional setting where the regression functions over a continuum of quantile levels are of interest. We propose a two-step penalization procedure, which accommodates stochastic integral based estimating equations and address the challenges due to the recursive nature of the procedure. We establish the uniform convergence rates for the proposed estimators, and investigate the properties on weak convergence and variable selection. We conduct numerical studies to confirm our theoretical findings and illustrate the practical utility of our proposals.

EMAIL: qi.zheng@louisville.edu

Testing for Marginal Effects in Quantile Regression

Huixia Wang*, The George Washington University
Ian McKeague, Columbia University
Min Qian, Columbia University

We develop a new marginal testing procedure to detect the presence of significant predictors associated with the quantiles (e.g., median) of a scalar response. The idea is to fit marginal quantile regression models based on each predictor separately, then select the predictor that minimizes the empirical quantile loss function and use the corresponding slope estimate as a test statistic. A resampling method is devised to calibrate this test statistic, which has non-regular limiting behavior due to the variable selection. Asymptotic validity of the procedure is established in a general quantile regression setting in which the marginal quantile regression models can be misspecified. Even though a fixed dimension is assumed to derive the asymptotic results, the proposed test is applicable and computationally feasible for large-dimensional predictors. The method is more flexible than existing

marginal screening test methods based on mean regression, and has the added advantage of being robust against outliers. The approach is illustrated using an application to an HIV drug resistance dataset.

EMAIL: judywang@gwu.edu

81. Diverse Evidence and Perspectives in Health Care Decisions

› Utility Maximizing Models of Medicare Supplemental Insurance Choices

Laura A. Hatfield*, Harvard Medical School
Jeannie Biniek, Harvard University
Melissa Favreault, Urban Institute
Michael Chernew, Harvard Medical School
Thomas McGuire, Harvard Medical School

Medicare households devote a much larger share of income to health care than non-retirees. Although Medicare covers the lion's share of health care expenses, substantial cost-sharing remains, and thus many beneficiaries elect supplemental coverage. Embedded within a large microsimulation model for health care spending among Medicare beneficiaries, we develop models to dynamically update beneficiaries' supplemental coverage. These models combine economic theory on utility maximization with multiple sources of information: beneficiary survey data, marginal distributions of coverage, and estimates of key decision-relevant parameters from the literature. One key challenge is calibrating the functional form of the utility function, specifically the interactions between personal characteristics and plan characteristics that form the basis of consumer decisions. We find that both utility maximization and simpler transition matrix approaches can capture the dynamic coverage choices of Medicare beneficiaries.

EMAIL: hatfield@hcp.med.harvard.edu

› Bayesian Hierarchical Regression and Variance-Function Modeling to Estimate the Inter-Rater Intraclass Correlation Coefficient in Assessments of Shared-Decision-Making

James O'Malley*, Geisel School of Medicine at Dartmouth
Paul J. Barr, Geisel School of Medicine at Dartmouth

I describe a novel methodological approach to assessing the interrater reliability of assessments of shared decision-making in a patient-physician clinical encounter. Two-raters each assess recordings of patient-physician encounters across three clinical sites using the Option5 shared-decision-making tool. The desired output is the interrater intraclass correlation coefficient (ICC) of the OPTION5 scores, accounting for heterogeneity of ratings between studies and the dependence of the between rater variance on the amount of shared decision-making. In this talk, I'll describe a Bayesian heteroscedastic hierarchical model that accomplishes these objectives and evaluate the impact of using this model on the estimated ICC. The resulting ICC will be shown to vary widely depending on whether the encounters being distinguished are restricted to the same site and to the amount of shared decision making in the encounter. Because current applied practice in shared decision often ignores these subtleties, this supports the introduction of standards regarding the definition of ICC and other measures of interrater reliability in shared decision-making.

EMAIL: Alistair.J.O'Malley@dartmouth.edu

› Integrating Patient Voice and Experience

Laura Lee Johnson*, U.S. Food and Drug Administration

Many methods can be used to estimate benefit and risk, and each method has drawbacks. Regardless of the method chosen, it is important to step back and understand if the data collected is a robust assessment of patient information and what information should be assessed. Additionally, it is not always feasible to give concrete information about evidence and uncertainty about unproven treatments with unknown side effects; how can this be conveyed and how can we still derive useful information? In particular how do we approach this in low literacy and low numeracy populations? We will describe sampling issues and examples of

the broad range of people to involve in this work when eliciting patients or caregivers considerations with respect to effectiveness and efficacy, safety, means of therapy implementation, duration of use and effect, other characteristics that may inform assessments, and patients' willingness (and unwillingness) to tolerate therapies. We will discuss mixed methods approaches to integrate qualitative and quantitative patient input and experience about the relevance of clinical endpoints to describe how patients feel, function, and survive.

EMAIL: laura.johnson@fda.hhs.gov

82. Variable Selection in Mixture of Regression

› A Hierarchical Hidden Markov Model Framework for Predicting Differential Methylation Profiles

Mayetri Gupta*, University of Glasgow
Tushar Ghosh, University of Glasgow
Neil Robertson, University of Glasgow
Peter Adams, University of Glasgow

DNA Methylation is an important epigenetic mechanism for controlling gene expression, silencing or genomic imprinting in living cells. High-throughput sequencing methods to study DNA methylation include sequencing of sodium bisulfite-treated DNA (BS-Seq). Several software tools for pre-processing and alignment of BS-seq data analysis have been published-however, methods for analyzing the methylation profiles, and detecting differentially methylated regions (DMRs) are relatively primitive, not taking specific dependence features of the data into account. Most current methods to detect DMRs rely on smoothing techniques that have high false discovery rates or are biased by experimental artefacts. We propose a novel method for predicting DMRs within a hierarchical Bayesian hidden Markov model framework, incorporating several levels of dependence between observations. Our method efficiently deals with nuisance parameters without leading to overwhelming analytical complexity and allows a principled way of building prior distributions based on partially known information, improving estimation of novel features. Our methods are illustrated through a study on human aging.

EMAIL: mayetri.gupta@glasgow.ac.uk

› Mixture of Generalized Linear Regression Models for Species-Rich Ecosystems

Frederic Mortier*, CIRAD, France

Understanding how climate change could impact population dynamics is of primary importance for species conservation. In species-rich ecosystems with many rare species, the small population sizes hinder a good fit of species-specific models. We propose a mixture of regression models with variable selection allowing the simultaneous clustering of species into groups according to vital rate information (recruitment, growth, and mortality) and the identification of group-specific explicative environmental variables. We illustrate the effectiveness of the method on data from a tropical rain forest in the Central African Republic and demonstrate the accuracy of the model in successfully reproducing stand dynamics and classifying tree species into well-differentiated groups with clear ecological interpretations.

EMAIL: fmortier@cirad.fr

› Mixture of Regression/Classification Trees

Emanuele Mazzola*, Dana-Farber Cancer Institute
Mahlet Tadesse, Georgetown University
Giovanni Parmigiani; Dana-Farber Cancer Institute

When analyzing high-dimensional datasets, it is often relevant to uncover homogeneous subgroups of individuals (cluster structures), and identify subgroup-specific variables associated with the responses. Assuming a latent or unobserved clustering structure, we propose a general method based on mixtures of regression or classification trees to identify relevant predictors associated to normally distributed (regression tree) or binary (classification tree) outcomes. We formulate the model in a Bayesian framework, allowing identification of homogeneous subgroups while searching for covariates that may have nonlinear association with the outcome. This is accomplished by alternating between the update of cluster allocations and the update of tree structures within the clusters at each MCMC iteration. We will illustrate the performance of the method using simulated and real datasets.

EMAIL: mazzola@jimmy.harvard.edu

83. Functional Data Analysis

Nonseparable Gaussian Stochastic Process: A Unified View and the Computational Strategy

Mengyang Gu*, Johns Hopkins University
Yanxun Xu, Johns Hopkins University
Barbara Engelhardt, Princeton University

We introduce a general framework of using Gaussian Stochastic Processes (GaSP) models for modeling multiple functional data with an ultra-fast and exact algorithm in large-scale problems and big data. GaSP models are widely used in modeling functional data, including emulation, interpolation, classification and uncertainty quantification. While GaSP is flexible from a modeling aspect, the major limitation arises in the evaluation of the likelihood, as it is at the order of $O(n ^ 3)$, where n is the number of sample size. We propose a general class of nonseparable GaSP models, in which the resulting computation can be done at the order of $O(n)$ without doing approximation, even when neither the covariance matrix nor precision matrix (the inverse of the covariance matrix) is sparse or low-rank. We show the popular linear regression and separable GaSP models are special cases of the proposed nonseparable methods. The advantages of the proposed nonseparable GaSP model are illustrated in an epigenetic application in which methylation levels are interpolated.

EMAIL: mgu6@jhu.edu

Longitudinal Dynamic Functional Regression

Md Nazmul Islam*#, North Carolina State University
Ana-Maria Staicu, North Carolina State University
Eric Van Heugten, North Carolina State University

We propose a statistical framework to study the dynamic association between scalar outcomes and functional predictor that evolves over time. The novelty is in the incorporation of time-varying functional effect and the proposal of a parsimonious approximation in the absence of prior information about the shape of functional coefficient. We introduce an efficient estimation procedure that has excellent numerical properties and allows to predict the full trajectories of outcome variable. We illustrate the ideas with extensive simulation study and an application to the lactating sow data, where the prediction is of primary concern; our method exhibits excellent numerical performance in terms of both prediction efficiency and computation time.

EMAIL: mnislam@ncsu.edu

A Joint Optimal Design for Functional Data with Application to Scheduling Ultrasound Scans

So Young Park*#, North Carolina State University
Luo Xiao, North Carolina State University
Jayson Wilbur, Metrum Research Group
Ana-Maria Staicu, North Carolina State University
N.L. Jumbe, Bill & Melinda Gates Foundation

We study a joint optimal design problem for sampling functional data. The goal is to find optimal time points for sampling functional data so that the underlying true function as well as a scalar outcome of interest can be accurately predicted. The problem is motivated by a fetal growth study, where the objective is to determine the optimal times to collect ultrasound measurements and the number of ultrasound measurements that are needed to recover fetal growth trajectories and to predict child birth outcomes. Under the frameworks of functional principal component analysis and functional linear models, we formulate the joint design into an optimization problem and the solution provides the optimal design points. We also propose a simple method for selecting the number of optimal sampling points. Performance of the proposed method is thoroughly investigated via a simulation study and by its application to the fetal ultrasound.

EMAIL: spark13@ncsu.edu

Significance Tests for Time-Varying Covariate Effect in Longitudinal Functional Data

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

We consider time-varying functional regression models to describe associations between longitudinal functional responses, where functions are observed at multiple instances (often

visit times) per subject for many subjects, and subject specific covariates. We develop inferential methods to assess the significance of the time-varying covariate effect. We propose a pseudo F- testing procedure that accounts for the complex error structure and is computationally efficient. Numerical studies confirm that the testing approach has the correct size and compares favorably with available competitors in terms of power. The methods are illustrated on a data application.

EMAIL: soh3@ncsu.edu

› Methodological Issues in the Functional Data Analysis of Actigraphy Data

Jordan S. Lundeen*, Augusta University
W. Vaughn McCall, Augusta University
Stephen W. Looney, Augusta University

This presentation examines several methodological issues we have encountered when using functional data analysis (FDA) to analyze actigraphy data. For example, we discuss and compare methods used for handling missing actigraphy data, and how to determine the optimal number of basis functions to use when applying FDA. Curves fit to actigraphy data must take on non-negative values, so we also discuss how to restrict FDA curves so that they have no negative values. The methods and issues we discuss are illustrated using actigraphy data from our study of the utility of a rest-activity biomarker to predict responsiveness to antidepressants. Among the biomarkers we consider are the acrophase (maximum) and bathyphase (minimum) of the actigraph-obtained activity levels during a given 24-hour period. We discuss the unique challenges we encountered in obtaining these summary measures of actigraphy data.

EMAIL: jolundeen@augusta.edu

› Quantile Functional Model

Hojin Yang*, University of Texas MD Anderson Cancer Center
Veera Baladandayuthapani, University of Texas MD Anderson Cancer Center
Jeffrey S. Morris, University of Texas MD Anderson Cancer Center

The aim of this paper is to develop quantile functional model framework accounting for entire quantile trajectories of the outcome given on the covariate for the subject. We develop an estimation procedure consisting of three stages as unified methodology. Our method first estimates empirical basis coefficients for each observed quantile curve, which characterizes individual monotone increasing quantile curve. Moreover, we use Markov chain Monte Carlo methods to obtain posterior samples for unknown parameters in the quantile functional model framework. Finally, we will acquire a quantity interested as some function of the entire quantile trajectory for each individual subject .Our simulation results and real data analysis show that our estimation methodology is useful and general approach.

EMAIL: hojiny0504@gmail.com

84. Computational Methods

› Enumerating Sets with Fallbacks Based on a Probability Model for a Kidney Paired Donation Program

Wen Wang*, University of Michigan
Mathieu Bray, University of Michigan
Peter X.K. Song, University of Michigan
John Kalbfleisch, University of Michigan

Kidney paired donation program (KPDP) is a partial solution to biological incompatibility preventing kidney transplants. A KPDP consists of non-directed donors (NDDs) and pairs, each of which comprises a candidate in need of a kidney transplant and her/his willing but incompatible donor. Potential transplants from NDDs or donors in pairs to compatible candidates in other pairs are determined by computer assessment and are subject to various uncertainties. A KPDP can be viewed as a directed graph with NDDs and pairs as vertices and potential transplants as edges, where a failure probability is attached to each vertex and edge. Transplants are carried out in the form of directed cycles among pairs and directed paths initiated by NDDs. Previous research shows that selecting disjoint subgraphs with a view to creating fallback options when failures occur generates more realized transplants than optimal selection of disjoint chains and cycles. In this study, we define such subgraphs, which are called locally relevant

subgraphs, and present an efficient algorithm to enumerate all LR subgraphs. Its computational efficiency is significantly better than the previous algorithms.

EMAIL: wangwen@umich.edu

› Optimization Methods for Generalized Linear Tensor Regression Models With Applications to fMRI Data

Dustin Pluta*, University of California, Irvine
Zhaoxia Yu, University of California, Irvine
Hernando Ombao, University of California, Irvine

Studies in neuroimaging require the analysis of complexly structured, high-dimensional data. Tensor regression offers a flexible approach to modeling the inherent structure in these data while also giving control of model complexity through tensor decompositions. In order to effectively apply tensor regression to complex datasets, it is necessary to understand the performance and limitations of current methods. This study evaluates the statistical and computational performance of generalized tensor regression models under different experimental settings, with and without regularization. The current method for fitting these models is adapted from the alternating least squares and iteratively reweighted least squares algorithms. Our simulation results suggest that adopting modern optimization methods can lead to improved model fit and reduced computation time. The effect of choice of rank for the tensor decomposition, which largely controls the model complexity, is also examined. Following the results of the simulation studies, a Poisson tensor regression model is applied to real data consisting of fMRI, genotype, and behavioral response data.

EMAIL: dpluta@uci.edu

› Divide-and-Conquer for Covariance Matrix Estimation

Gautam Sabnis*, Florida State University
Debdeep Pati, Florida State University
Barbara Engelhardt, Princeton University
Natesh Pillai, Harvard University

We propose a Divide-and-Conquer (D&C) strategy, a parallel framework, to accelerate posterior inference for high-dimensional covariance matrix estimation using Bayesian latent factor models. D&C distributes the task of high-dimensional covariance matrix estimation to multiple machines, solves each sub-problem via modeling a latent factor model, and then combines the sub-problem estimates to produce a global estimate of the covariance matrix. The existing D&C methods almost exclusively focus on tackling large n problems when the data points are independent and identically distributed. Our approach is different from current methods in that it leaves the number of samples n unchanged but, instead, splits the original dimension p into smaller sub-dimensions. We propose a novel hierarchical structure on the latent factors that allows flexible dependence across estimates, obtained from different machines, while still maintaining the computational efficiency. The approach is readily parallelizable and is shown to have computational efficiency of several orders of magnitude compared to fitting a full factor model.

EMAIL: gss12b@my.fsu.edu

› A Modified Conditional Metropolis-Hastings Sampler Under Two Generalized Strategies

Jianan Hui*, University of California, Riverside
James Flegal, University of California, Riverside
Alicia Johnson, Macalester College

A modified conditional Metropolis–Hastings sampler for general state spaces is investigated under two generalized strategies. Under specified conditions, we show that the generalization of the modified conditional Metropolis–Hastings sampler can also lead to substantial gains in statistical efficiency while maintaining the overall quality of convergence. Results are illustrated in both simulated and real data. For the simulated data settings, we use a toy bivariate Normal model and a Bayesian version of the random effects model. In order to illustrate its utility in high-dimensional simulations, we consider a dynamic space-time model on weather station data.

EMAIL: jhui003@ucr.edu

Contradiction or Contribution? Introducing StatTag, A New Tool to Conduct Reproducible Research using Microsoft Word

Leah J. Welty*, Northwestern University
Luke V. Rasmussen, Northwestern University
Abigail S. Baldridge, Northwestern University
Eric W. Whitley, Independent Consultant

This talk will introduce StatTag, a free plug-in for conducting reproducible research using Microsoft Word and Stata, SAS, and R. StatTag was developed to address a critical need in the research community: there were no broadly accessible tools to integrate document preparation in Word with statistical code and results. Popular tools such as knitR and Markdown use plan text editors for document preparation. Despite the merits of these programs, Microsoft Word is ubiquitous for manuscript preparation in many fields, such as medicine, in which conducting reproducible research is increasingly important. Furthermore, current tools are one-directional: no downstream changes to the rendered RTF/Word document are reflected in the source code. We developed StatTag to fill this void. StatTag provides an interface to edit statistical code directly from Word, and allows users to embed statistical output from that code (estimates, tables, figures) within Word. Output can be updated in one-click with a behind-the-scenes call to the statistical program. With StatTag, modification of a dataset or analysis no longer entails transcribing results into Word.

EMAIL: lwelty@northwestern.edu

A General Framework for the Regression Analysis of Pooled Biomarker Assessments

Yan Liu*#, Clemson University
Christopher McMahan, Clemson University
Colin Gallagher, Clemson University

As a cost efficient data collection mechanism, the process of assaying pooled biospecimens is becoming increasingly common in epidemiological research; e.g. pooling has been proposed for evaluating the diagnostic efficacy of biological markers (biomarkers). To this end, several authors have proposed techniques to analyze continuous pooled biomarker assessments. Regretfully, most of these techniques proceed under restrictive assumptions, are unable to account for the effects of measurement error, and fail to control for confounding variables. Consequently, a general Monte Carlo maximum likelihood based procedure is presented. The proposed approach allows for the regression analysis of pooled data under practically all parametric models and can be used to directly account for the effects of measurement error. Through simulation, it is shown that the proposed approach can accurately and efficiently estimate all unknown parameters. This new methodology is further illustrated using monocyte chemotactic protein-1 data collected by the Collaborative Perinatal Project in an effort to assess the relationship between this chemokine and the risk of miscarriage.

EMAIL: yan5@g.clemson.edu

85. Dynamic Treatment Regimens

Cumulative Incidence Regression for Dynamic Treatment Regimes

Ling-Wan Chen*, University of Pittsburgh
Idil Yavuz, Dokuz Eylul University
Yu Cheng, University of Pittsburgh
Abdus S. Wahed, University of Pittsburgh

Recently dynamic treatment regimes (DTRs) have drawn considerable attention, and two-stage randomization is often used to gather data for making inference on DTRs. Meanwhile, practitioners become aware of competing-risk censoring, where subjects are exposed to multiple failures, and the target event may not be observed due to the occurrence of competing events. In this paper, we focus on regression analysis of DTRs from a two-stage randomized trial for competing-risk censored outcomes based on cumulative incidence functions (CIFs). Even though there are extensive works on the regression problem for DTRs, no research has been done on modeling the CIF when a two-stage randomization has been carried out. We extend existing CIF regression methods to model covariates effects for DTRs. Asymptotic properties are established for our proposed estimators. The models can be implemented using standard software by an augmented-data approximation. We show the improvement of

our proposed methods by simulation, and illustrate its practical utility through an analysis of a two-stage randomized neuroblastoma study, where disease progression is subject to competing-risk censoring by death.

EMAIL: lic76@pitt.edu

› A Principled Policy-Optimized Bayesian Nonparametric Formulation of Periodontal Recall Intervals

Qian Guan*#, North Carolina State University
Brian Reich, North Carolina State University
Eric Laber, North Carolina State University
Dipankar Bandyopadhyay, Virginia Commonwealth University

Tooth loss from periodontal disease is a major public health burden in the United States. Standard clinical practice is to recommend a dental visit every six months; however, this practice is not data-driven, and poor dental outcomes and increasing dental insurance premiums indicate room for improvement. We consider a tailored approach that recommends treatments if, when, and to whom they are needed thereby simultaneously reducing periodontal disease and resource expenditures. We formalize this tailoring as dynamic treatment regime which comprises a sequence of decision rules. The dynamics of periodontal health, visit frequency, and patient compliance are complex yet the estimated optimal regime must be interpretable to domain experts. We combine flexible non-parametric Bayesian dynamics modeling with policy-search algorithms to estimate the optimal dynamic treatment regime with an interpretable class of regimes. Both simulation experiments and application to a rich database of electronic dental records from the HealthPartners confirms the effectiveness of our proposed methodology relative to existing methods.

EMAIL: qguan2@ncsu.edu

› Using Multiple Comparisons with the Best to Identify the Best Set of Dynamic Treatment Regimens from a SMART with a Survival Outcome

Qui Tran*, University of Michigan
Kelley M. Kidwell, University of Michigan
Alexander Tsodikov, University of Michigan

Given that multiple first- and second-line treatment options are available for many diseases, clinical decision making processes require algorithms to find the optimal decision rules to assign treatment combinations that result in the best response over the course of treatment, instead of stage-specific optimization. A sequential multiple assignment randomized trial (SMART) is one such type of trial that has been designed to assess the entire treatment regimen. In such a trial, second-line treatment is assigned based on a patient's outcome to the first-line treatment. These trials generally include many dynamic treatment regimens (e.g. 8) leading to a multiple comparisons problem. We have developed methods to analyze data from a SMART with a survival outcome addressing the multiple comparison problem using multiple-comparisons-with-the-best (MCB). Simulation studies are performed to demonstrate the effectiveness of the proposed methods.

EMAIL: quitran@umich.edu

› Incorporating Patient Preferences into Estimation of Optimal Individualized Treatment Rules

Emily L. Butler-Bente*, University of North Carolina, Chapel Hill
Eric B. Laber, North Carolina State University
Sonia M. Davis, University of North Carolina, Chapel Hill
Michael R. Kosorok, University of North Carolina, Chapel Hill

Individualized treatment rules operationalize precision medicine as a map from patient information to a recommended treatment and seek to maximize the mean of a scalar outcome. However, in settings with multiple outcomes, choosing a scalar composite outcome to define optimality is difficult. Furthermore, due to heterogeneity across patient preferences for these outcomes, it may not be possible to construct a single composite outcome that leads to high-quality treatment recommendations for all patients. We simultaneously estimate the optimal individualized treatment rule for all composite outcomes representable as a convex combination of the outcomes. For each patient, we use a preference elicitation questionnaire and item response theory to derive the posterior distribution over patient preferences and derive an estimator of an optimal individualized treatment rule tailored to patient preferences. We prove that as the number of sub-

jects and items on the questionnaire diverge, our estimator is consistent for an oracle optimal individualized treatment regime. We illustrate the proposed method using data from a clinical trial on antipsychotic medications for schizophrenia.

EMAIL: emily.lynn.butler@gmail.com

› Discovering Treatment Effect Heterogeneity Through Post-Treatment Variables with Application to the Effect of Class Size on Math Scores

Ashkan Ertefaie*, University of Rochester
Jesse Y. Hsu, University of Pennsylvania
Dylan S. Small, University of Pennsylvania

Class-size reduction has been supported by many education policy makers. However, because of limited resources, a crucial question is, are there types of students who benefit more from small classes, i.e., are there effect modifiers for the benefit of small classes? We use data from the Tennessee STAR study, a large class size randomized experiment to address this question. In the Tennessee STAR study, relatively few potential effect modifiers were measured at baseline but many potential effect modifiers were measured after baseline. While treatment effect modification based on pretreatment variables in a randomized trial can be assessed using standard regression, for post-treatment variables such regression approaches would only be valid under a strong sequential ignorability assumption. In this paper, we develop a method for studying effect modification based on post-treatment variables that does not rely on this strong sequential ignorability assumption. We provide evidence that students who are not motivated benefit more from small classes than students who are motivated.

EMAIL: ashkan_ertefaie@urmc.rochester.edu

› Tree-Based Reinforcement Learning for Estimating Optimal Dynamic Treatment Regimes

Yebin Tao*, Eli Lilly and Company
Lu Wang, University of Michigan
Daniel Almirall, University of Michigan

Dynamic treatment regimes (DTRs) are sequences of treatment decision rules, in which treatment is adapted over time in response to the changing course of an individual. 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 the 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. We illustrate our method in a case study to identify dynamic substance abuse treatment regimes for adolescents.

EMAIL: yebintao@umich.edu

86. Semi-Parametric and Non-Parametric Models

› Semiparametric Model and Inference for Spontaneous Abortion Data with a Cured Proportion and Biased Sampling

Jin Piao*, University of Texas Health Science Center at Houston
Jing Ning, University of Texas MD Anderson Cancer Center
Christina Chambers, University of California, San Diego
Ronghui Xu, University of California, San Diego

Understanding the risk and safety of using medications in a woman during her pregnancy will help both clinicians and pregnant women to make better treatment decisions. However, utilizing spontaneous abortion (SAB) data collected in observational studies poses two major challenges. First, the data from the observation cohort are not random samples of the target population due to the sampling mechanism. Pregnant women with early SAB are more likely to be excluded from the cohort. Second, the observed data are heterogeneous and contain a cured proportion. In this article, we consider semiparametric models to simultaneously estimate the probability of being cured and the distribution of time to SAB for the uncured subgroup. We appropriately adjust the sampling bias in the likelihood function and develop an expectation-maximization algorithm to overcome

the computational challenge. We apply the empirical process theory to prove the consistency and asymptotic normality of the estimators. We examine the finite sample performance of the proposed estimators in simulation studies and illustrate the proposed method through an application to SAB data.

EMAIL: jin.piao@uth.tmc.edu

› Learning Semiparametric Regression with Missing Covariates Using Gaussian Processes Models

Abhishek Bishoyi*, University of Connecticut
Dipak K. Dey, University of Connecticut
Xiaojing Wang, University of Connecticut

In this paper, we consider a semiparametric regression in the presence of missing covariates for nonparametric components under Bayesian framework. As known, Gaussian processes are a popular tool in nonparametric regression because of their flexibility and the fact that much of the ensuing computation is parametric Gaussian computation. In addition, the mean function of a Gaussian process can be modeled by a parametric function. But the most frequently used covariance functions of a Gaussian process will not be well defined in the absence of covariates. We propose an imputation method to solve this issue and perform our analysis using Bayesian inference, where we specify the objective priors on the parameters of Gaussian process models. We have shown that under mild conditions, such objective priors assigned would yield proper posterior, which has successfully extended the results of Berger et al. (2001) and Ren et al. (2012) into missing data framework. Several simulations are conducted to illustrate effectiveness of our proposed method and further, our method is exemplified through Langmuir equation, commonly used in pharmacokinetic models.

EMAIL: abhishek.bishoyi@uconn.edu

› Competing Risks Regression Models Based on Pseudo Risk Sets

Anna Bellach*, University of Copenhagen
Jason P. Fine, University of North Carolina, Chapel Hill
Michael R. Kosorok, University of North Carolina, Chapel Hill

We present a direct and general extension of the Fine-Gray model. Our new regression model is applicable to a broad class of semiparametric transformation models and targets the subdistribution hazard. It accommodates for time dependent covariates and allows for clinically relevant extensions such as the setting of recurrent events with competing terminal events, where we target the marginal mean intensity. The theoretical background of the proposed method is provided including asymptotic properties of the estimators, variance estimator, simulation studies, model selection and several applications to clinical study data. It is a common mistake in practice that competing terminal events are ignored or treated as censorings, which can lead to a biased estimation. Comparing our results for a bladder cancer data set to results from other methods, we underline the importance of accounting correctly for such competing terminal events.

EMAIL: annabella@sund.ku.dk

› Semiparametric Efficient Approach for a Linear Transformation Model with Errors-in-Covariates

DongHyuk Lee*, Texas A&M University
Yanyuan Ma, The Pennsylvania State University
Samiran Sinha, Texas A&M University

We propose a semiparametric efficient estimator for the linear transformation model when the time-to-events are subject to right censoring and covariates are measured with errors. The proposed method produces consistent estimators even when the assumed models for the true unobserved covariates are misspecified. We derive the asymptotic properties of the estimators, and the operating characteristics of the proposed method are assessed via finite sample simulation studies. Finally, we apply the proposed method to analyze a real data set.

EMAIL: dhyuklee@stat.tamu.edu

› A Semiparametric Inference of Geometric Mean with Detection Limit

Bokai Wang*, University of Rochester
Jing Peng, University of Rochester
Changyong Feng, University of Rochester

In biomedical research, geometric means are widely used to estimate and compare population geometric means of

non-negative-valued outcomes between different groups. In many studies, due to the limited ability to measure devices, quantities of interest are not always measured, a phenomenon known as detection limit. A common approach in biomedical research is to replace missing values by small positive constants and proceed with inference based on the imputed data. However, no work has been carried out to date to study potential effects of this imputation method on inference about the population geometric mean. In this paper, we take an in-depth look at this issue. Our findings show that this intuitive and innocent-looking strategy can actually change inference in a drastic way, making results uninterpretable, even if the detection limit is very small. We also propose a semiparametric approach to estimate the ratio of geometric means without imputing the missing data. Results show that our approach has significant implications for biomedical research, promoting an urgent call to put a stop to the use of the imputation method and to switch to the proposed approach.

EMAIL: bokai_wang@urmc.rochester.edu

› Moment-Based Semiparametric Regression Analysis of Pooled Biomarker Assessments

Juexin Lin*, University of South Carolina
Dewei Wang, University of South Carolina

In epidemiological research, the laboratory tests of evaluating biomarkers for disease detecting are often prohibitively expensive and time-consuming. For the cost and time effective purpose, pooling has been proposed. The research analyzing pooled biomarker assessments is commonly conducted under parametric assumptions. To relax such assumptions, we proposed a general semiparametric regression framework via moment matching that allows for the incorporation of individual-level covariates. The asymptotic properties of our estimators are derived and presented. We investigate the finite sample performance of our proposed method through simulation as well as practical diabetes data from NHANES. An interesting founding is that the estimates carried out from pooled strategy could be more accurate than those from testing individuals separately.

EMAIL: juexin@email.sc.edu

› A Semiparametric Joint Survival Model with a Time-Dependent Cure Process

Sophie Yu-Pu Chen*#, University of Michigan
Alexander Tsodikov, University of Michigan

Cure models have been applied to time-to-event data when a proportion of patients is not at risk following treatment. Current work on cure models focuses on the class of static models, where the cure status is assumed determined at the beginning of the follow up. In practice, patients often receive treatments, or intermediate events are observed during the follow up. In this case it is natural to expect the chance of cure to change in response to dynamic factors. To account this, we propose a joint dynamic model for the cure process and a terminal event. Two separate baseline hazards are estimated nonparametrically in the model to allow for different time scales for the time to cure and time to failure processes. An EM algorithm is developed to estimate the two infinite dimensional parameters. Covariates are modeled parametrically and their effects are estimated using the profile likelihood. Large-sample properties are obtained. Simulation studies are presented to illustrate the finite-sample properties. The proposed model is applied to study the effect of secondary cancer on survival following a primary prostate cancer diagnosis using data from SEER program.

EMAIL: yupuchen@umich.edu

87. Clustered Data Methods

› Outcome-Dependent Sampling in Cluster-Correlated Data Settings with Application to Hospital Profiling

Glen W. McGee*, Harvard University
Jonathan S. Schildcrout, Vanderbilt University
Sharon-Lise T. Normand, Harvard University
Sebastien Haneuse, Harvard University

Hospital readmission is a key marker of healthcare quality used by the Centers for Medicare and Medicaid Services (CMS) to determine hospital reimbursement rates. Analyses of readmission are based on a logistic-normal generalized linear mixed model (LN-GLMM) that permits estimation of hospital-specific measures while adjusting for case-mix

differences. Recently a bill was introduced to Congress that would require CMS to include currently unobserved measures of socioeconomic status in their case-mix adjustment without further burden to hospitals. We propose that detailed socioeconomic data be collected on a sub-sample of patients via a novel outcome-dependent sampling scheme: the cluster-stratified case-control design. Towards valid estimation and inference for both fixed and random effects components of an LN-GLMM, we propose methods based on (i) inverse-probability weighting, (ii) observed sample likelihood, and (iii) multiple imputation. Methods are motivated by data on 34,694 Medicare beneficiaries hospitalized between 2011-13 with a diagnosis of acute myocardial infarction. Small-sample operating characteristics and design considerations are evaluated via simulation.

EMAIL: glenmcgee@g.harvard.edu

› Improved Model Fitting Procedures for Multiple Membership Data Structures in Multilevel Models

Tugba Akkaya-Hocagil*, State University of New York at Albany
Recai M. Yucel, State University of New York at Albany

Model-fitting techniques for multilevel models have been extensively treated from both frequentist and Bayesian perspectives (e.g. iterative generalized least-squares, EM-based or MCMC-based algorithms). We consider these model-fitting paradigms the data structure is not purely hierarchical i.e. lowest level observational units to belong to multiple higher level clustering factors such as students attending multiple schools. We extend well-established computationally-efficient algorithms operating under linear mixed-effects models. These algorithms operate under a combination of Fisher's scoring and EM algorithms for finding maximum likelihood estimates as well as MCMC-based algorithms that provide excellent mixing properties. We illustrate these techniques through a comprehensive simulation study.

EMAIL: akkaya.tugba@gmail.com

› A New Statistical Approach for Few Clusters with Cluster-Level Heterogeneity

Ruofei Du*, University of New Mexico Comprehensive Cancer Center
Ji-Hyun Lee, University of New Mexico Comprehensive Cancer Center

In order to have reliable statistical analysis for cluster data, a small number of clusters has been a concern. Common standard analysis methods (e.g. linear mixed-effects modeling) are based on large number of clusters; however, it has long been realized this condition is hardly satisfied in many real practices. When insufficient number of clusters being used, in addition to lack of representation of the study population of clusters, one or two outlier clusters (i.e. heterogeneity at cluster level) could have significant impact on the statistical analysis. We propose a weighted delete-one-cluster Jackknife based framework to balance the influence from each cluster, aiming to achieve improved statistical analysis. We show, via simulations, the proposed framework has good operating characteristics with respect to the mean squared error of estimate, and statistical test power. The new algorithm is also applied to two real datasets from school-based surveys. It provides higher probability of drawing the same conclusion between using randomly chosen small number of schools (i.e. clusters), and the entire large number of schools.

EMAIL: rdu@salud.unm.edu

› Clusterability Evaluation: A Pre-Requisite for Clustering

Margareta Ackerman, San Jose State University
Andreas Adolfsson, Florida State University
Naomi C. Brownstein*, Florida State University

Clustering is an essential data mining tool that aims to discover inherent cluster structure in data. As such, the study of clusterability, which evaluates whether data possesses such structure, is an integral part of cluster analysis. Yet, despite their central role in the theory and application of clustering, there are fewer methods to evaluate the clusterability of a dataset, and the methods that do exist are spread across multiple disciplines and use different nomenclature. We synthesize a comprehensive collection of methods for

clusterability evaluation. Through extensive simulations and application on real datasets, we compare methods for validity, power, and computational complexity. Finally, we provide recommendations to decide which method to apply in practice for determining whether or not to cluster a given dataset.

EMAIL: naomi.brownstein@med.fsu.edu

› A Bayesian Approach to Zero-Inflated Clustered Count Data with Dispersions

Hyoyoung Choo-Wosoba*, National Cancer Institute, National Institutes of Health
Jeremy Gaskins, University of Louisville
Steven Levy, University of Iowa
Somnath Datta, University of Florida

We present a novel Bayesian approach for analyzing zero-inflated clustered count data with dispersion. In order to deal with such data, we combine the Conway-Maxwell-Poisson distribution which allows both over- and under-dispersion with a hurdle component for excess zeros and random effects for clustering. We propose an efficient Markov chain Monte Carlo sampling scheme to obtain posterior inference from our model. Through simulation studies, we compare our hurdle CMP model with a hurdle Poisson model to demonstrate the effectiveness of our CMP approach. Furthermore, we apply our model to analyze a dataset containing information on the number and types of carious lesions on each tooth in a population of 9 year olds from the Iowa Fluoride Study, which is an ongoing longitudinal study on a cohort of Iowa children that began in 1991.

EMAIL: hyoyoung.choo-wosoba@nih.gov

› A Semiparametric Method for Clustering Mixed-Type Data

Alexander H. Foss*, University at Buffalo
Marianthi Markatou, University at Buffalo

There are hundreds, if not thousands, of methods for cluster analysis in the computer science and statistics literature. Despite this large and growing body of methods, clustering mixed-type data (both continuous and categorical variables) remains a challenging task. We briefly review methods for clustering mixed-type data, highlighting deficiencies that broadly affect many of the most popular existing approaches. We propose a novel semiparametric method based on mixture models that addresses these issues. Algorithms for fitting our model are compatible with a map-reduce framework, and we discuss the implementation of our clustering method in Hadoop for analyzing very large data sets on distributed file storage systems.

EMAIL: ahfoss@buffalo.edu

88. Bringing Adaptive Trails Into the Real World

› Parametric Dose Standardization for Two-Agent Phase I-II Trials with Ordinal Efficacy and Toxicity

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

A Bayesian methodology is presented for jointly optimizing the doses of a two agent combination in phase I-II trials based on bivariate ordinal efficacy and toxicity. For each marginal distribution, a generalized continuation ratio dose-outcome model is assumed, with each agent's dose parametrically standardized in the linear term to accommodate possibly complex dose-outcome relationships. Elicited numerical utilities of the elementary outcomes are used to compute posterior mean utilities of all dose pairs, which are used as decision criteria. Outcome-adaptive randomization among dose pairs close to optimal is used to reduce the risk of getting stuck at a suboptimal dose pair. An extensive simulation study shows that the proposed methodology is robust, and it compares favorably with several alternative approaches, including designs based on conventional models with multiplicative dose-dose interactions.

EMAIL: peterthall6775@gmail.com

› Bayesian and Frequentist Adaptive Designs: Experience from the World of Medical Devices

Gregory Campbell*, GCStat Consulting

The wealth of experience on the use of Bayesian statistics in medical device clinical trials has been most helpful in the development of adaptive design principles. In 2010, FDA issued a

guidance on Bayesian clinical trials for medical devices following a decade of experience. Principles included 1) the importance of detailed planning and the pre-specification of the design; 2) a firm understanding of the operating characteristics of the design; and 3) the importance of simulation. This experience has been most helpful in the development of an FDA guidance on adaptive designs for medical device clinical studies, finalized in July, 2016. An additional principal concerns the minimization of bias, especially operational bias. Since most sample size calculations for fixed designs are often based on at best questionable assumptions, adaptive designs have been dramatically underutilized. While adaptive designs require intricate planning and often the difficult task of anticipated regret supposing the trial does fail, such designs can greatly improve both the efficient use of trial resources and the ultimate probability of trial success.

EMAIL: patgregcampbell@verizon.net

› Avoiding Bias in Longitudinal Internal Pilot Studies

Xinrui Zhang*, University of Florida
Yueh-Yun Chi, University of Florida

When planning a longitudinal study with Gaussian outcomes, accurate specification of the variances and correlations is required to select an appropriate sample size. Underspecifying the variances leads to a sample size that is inadequate to detect a meaningful effect, while overspecifying the variances results in an unnecessarily large sample size. Both place study participants in unwarranted risk. An internal pilot design protects against covariance misspecification by using a fraction of the observed data to re-estimate the covariance matrix and adjust the sample size. We derive an approximate distribution of the univariate approach to repeated measures (UNIREP) test statistic with an account of the randomness in the final sample size. We apply a bounding approach to modify the critical value and ensure the maximum Type I error rate is at or below the target. The bounding approach is extended to longitudinal data with serial correlation of various decay rates. Enumeration and simulation results demonstrate the accuracy of the proposed methods in controlling the Type I error rate while maintaining the benefits of an internal pilot design in preserving power.

EMAIL: xinrui@ufl.edu

› Increasing the Practicality of Innovative Trial Design

Christopher S. Coffey*, University of Iowa

Adaptive designs allow reviewing accumulating information during an ongoing clinical trial to possibly modify trial characteristics. To preserve study validity, changes should be based on pre-specified decision rules. This often requires properly designed simulation studies, and has led to an industry movement towards in-house teams responsible for planning and conducting such simulations. Greater barriers exist for implementing similar infrastructure within the academic clinical trials environment. Consequently, there is a growing divide between the feasibility of conducting adaptive designs in industry compared to academia. General acceptance of adaptive designs requires increasing their use across all types of clinical trials. Thus, infrastructure building efforts are needed within the academic clinical trials environment. In this presentation, I provide a summary of the NINDS-funded Network for Excellence in Neuroscience Clinical Trials (NeuroNEXT) experience to date and illustrate how this infrastructure increases the feasibility for using more novel trial designs in an academic trials setting.

EMAIL: christopher-coffey@uiowa.edu

89. Analyzing Clinical Trial Results in the Presence of Missing Data: Anything Better Than Imputation?

› Seven Kinds of Missing Data

Thomas J. Permutt*, U.S. Food and Drug Administration

In a sample survey there is approximately one kind of missing data: a piece of information about a subject is unknown, but well-defined, it exists, and has to be ascertained to produce an estimate of a meaningful aggregate. In a clinical trial, irregularities of many different kinds may occur. Subjects may die; take other drugs; drop out due to adverse events or lack of efficacy; be lost to follow-up for reasons that are not ascertained. Some of these events may absolutely prevent observation of the primary outcome; others may allow observation of a value that might be considered tainted. Some may represent protocol violations, but others may be foreseen in the protocol. It is

inconceivable that all such irregularities should be handled the same way as missing data in surveys. Some must be treated as outcomes in themselves, and others may advantageously be so treated. Sometimes a straightforward analysis of the tainted values may suffice, but careful thought should be given to its interpretation. Sometimes the aggregate measure can be defined in such a way that no or few data are missing. Protocols should discuss all these aspects.

EMAIL: ashli.walker@fda.hhs.gov

› Can We Avoid Counterfactuals by Using Missingness Itself in the Endpoint?

Michael P. O'Kelly*, QuintilesIMS

Estimands that "define away" missingness, for example seeking an estimate only in tolerators or only in completers, do not generally answer a clinically useful question. Estimands that include outcomes after randomized treatment is discontinued could reduce structural missingness, but lose the advantage of a purely randomized analysis. Could we include missingness in our endpoint, without in some way "telling a story" about the missing data, i.e. without modelling what might have been, had the missing data been observed? Options here include treating subjects with missing outcomes as failures; or using some form of trimmed mean that implies a "bad" outcome when the outcome is missing. Examples will show strengths and weaknesses of these approaches. Rubin's suggestions for making causal inference using potential outcomes are described as a view that could help to undo, or at least loosen, the Gordian knot of missing data in clinical trials.

EMAIL: mokelly@quintiles.com

› Quantifying the Properties of Statistical Methods Handling Missing Data in Clinical Trials - What Can We Learn?

Elena Polverejan*, Janssen R&D

There are many challenges in the selection, for a clinical trial, of the primary estimand and corresponding analysis methods in the presence of missing response information. Each trial is unique, with its own characteristics such as objective, indication, population, design, efficacy response, likelihood of subjects remaining on treatment or in the trial, type of potential trial discontinuations and so forth. There are also many statistical methods to choose from, all relying on different assumptions. Some methods impute what has not been observed and some treat missingness as an outcome. This presentation uses a clinical trial simulation exercise to compare various statistical methods, with or without imputation, under several assumptions for the trial. Simulation metrics such as the estimated power and Type I error rate, mean and standard error for the treatment difference versus control, etc. allow a quantitative comparison of the performance of different methods. The talk will show how the simulation results play a critical role towards an informed selection of the primary analysis method and additional sensitivity analyses.

EMAIL: EPolvere@its.jnj.com

90. Large-Scale Spatio-Temporal Decision Problems

› Doubly Robust Estimation in Observational Studies with Partial Interference

Lan Liu, University of Minnesota
Michael Hudgens*, University of North Carolina, Chapel Hill
Bradley Saul, University of North Carolina, Chapel Hill

Interference occurs when the treatment received by one individual affects the outcome of another individual. Partial interference refers to the particular setting wherein individuals can be partitioned into clusters such that interference may occur within clusters but not between clusters. In observational studies where individuals are not randomly assigned treatment, inverse probability weighted (IPW) estimators have been proposed under the partial interference assumption. However, the validity of these IPW estimators depends on correct specification of the propensity score model. Alternatively, treatment effects in the presence of partial interference can be estimated using outcome regression estimators which depend on correct specification of the outcome model. In this talk we consider doubly robust estimators, which are consistent if either the propensity score model or the outcome regression model (but not necessarily both) is correctly specified. Empirical results are presented demonstrating the doubly robust-

ness of the proposed estimators. The different estimators are illustrated using data from a large cholera vaccine study.

EMAIL: mhudgens@bios.unc.edu

› A Spatial Data Fusion Approach for Causal Inference

Alexandra Larsen, North Carolina State University
Ana Rappold, U.S. Environmental Protection Agency
Brian Reich*, North Carolina State University

Causal inference is challenging for spatially-dependent environmental data because of the lack of randomization and true replication. Mathematical models offer a potential solution because the user can simulate the process under different scenarios to explore intervention effects. However, mathematical models are often biased and parameterized at too coarse of a scale to replicate extreme events. In this talk we propose a new spatial approach to combine an ensemble of numerical models with observational data to estimate causal effects. We apply the new method to estimate the contribution of fire smoke to overall air pollution in different regions of the US.

EMAIL: brian_reich@ncsu.edu

› Spatial Spread of the West Africa Ebola Epidemic at Two Scales

John Drake*, University of Georgia

The 2013-2015 epidemic of Ebola virus disease in West Africa is the largest documented, affecting >20,000 persons in three countries. Achieving control of localized outbreaks during Ebola containment and effective preemptive response to future emerging diseases will be enhanced by understanding the role of geography in promoting or inhibiting human-to-human transmission and evaluating the relative probability of alternative spatial epidemic paths. Using generalized gravity models we characterized the spatial network over which Ebola virus spread and we estimated the effects of geographic covariates on transmission rate during peak spread. We also introduce a new graphical device for visualizing dynamic probabilities on graphs, the probability path plot. As a tool for communicating with policymakers and non-technical audiences, the probability path plot concisely represents how risk changes based on both local and non-local context. These findings highlight the importance of integrated geography to epidemic expansion and containment and may contribute to identifying both the most vulnerable unaffected areas and locations of maximum intervention value.

EMAIL: jdrake@uga.edu

91. Use Of Historical Information for Evidence Synthesis and Decision Making

› Using Historical Patient Data to Estimate Population Treatment Effects

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

With increasing attention being paid to the relevance of studies for real-world practice, there is growing interest in external validity and assessing whether the results seen in randomized trials would hold in target populations. While randomized trials yield unbiased estimates of the effects of interventions in the sample in the trial, they do not necessarily inform about what the effects would be in some other, potentially somewhat different, population. Relatively little statistical work has been done developing methods to assess or enhance the external validity of randomized trial results. In addition, data sources such as information on historical patients provide information on broad target populations, including individuals who may not agree to participate in a trial. This talk will discuss design and analysis methods for assessing and increasing external validity. Underlying assumptions, performance in simulations, and limitations will be discussed. Implications for how future studies should be designed to enhance generalizability will also be discussed.

EMAIL: estuart@jhsph.edu

› Leveraging Historical Controls Using Multisource Adaptive Design

Brian P. Hobbs*, University of Texas MD Anderson Cancer Center
Nan Chen, University of Texas MD Anderson Cancer Center
Bradley P. Carlin, University of Minnesota

Health care practice evolves through the continual endeavor to enhance current strategies through clinical study. Beneficial

therapeutic strategies are established through a gradual process devised to define the safety and efficacy profiles of new strategies over a sequence of clinical trials. This system produces redundancies, whereby similar treatment strategies are replicated, either as experimental or comparator standard-of-care therapies, across development phases and multiple studies. This article describes a collection of web-based statistical tools hosted by MD Anderson Cancer Center that enable investigators to incorporate historical control data into analysis of randomized clinical trials using Bayesian hierarchical modeling as well as implement adaptive designs using the method described in Hobbs et al. (2013). By balancing posterior effective sample sizes among the study arms, the adaptive design attempts to maximize power on the basis of interim posterior estimates of bias. With balanced allocation guided by dynamic Bayesian hierarchical modeling, the design offers the potential to enhance efficiency while limiting bias and controlling average type I error.

EMAIL: bphobbs@mdanderson.org

› Combination Dose Finding in Oncology: A Co-Data Approach

Satrajit Roychoudhury*, Novartis Pharmaceuticals Corporation

In recent years, there is a growing interest in the development of combinations of new investigational drugs for treating cancer. At the early phase of development of new drug multiple phase I trials with different combination partners are conducted in parallel. The primary objective of these trials is to find appropriate combination(s) and dose(s) for further development. Selecting best potential combination(s) requires good knowledge of safety and efficacy profiles of multiple combination agents. However, this can be challenging due to small trial size and heterogeneity between trials. We propose a model based evidence synthesis approach to characterize safety and efficacy of a new drug in combination with different combination partners. The proposed approach synthesizes all available single agent and combination data from different Phase I trials with multiple drug combinations to increase the precision in statistical inference. Therefore, this leads to a robust decision-making in terms of optimal combination selection.

EMAIL: satrajit.roychoudhury@novartis.com

92. Use of Real-World Evidence for Regulatory-Decision Making: Statistical Considerations and Beyond

› Incorporating Real World Evidence for Regulatory Decision Making: Challenges and Opportunities

Lilly Q. Yue*, U.S. Food and Drug Administration
Nelson T. Lu, U.S. Food and Drug Administration
Yunling Xu, U.S. Food and Drug Administration

In this era of "Big Data," there are many sources of real world healthcare data that could be potentially leveraged in the clinical studies in the regulatory settings. While such large quantities of data reflect real world clinical practice and could potentially be used to reduce the cost for bringing medical products to the market, challenges arise concerning transforming the real world data into valid scientific evidence and using such evidence in the regulatory decision making. This presentation will discuss the opportunities and challenges from statistical and regulatory perspectives with examples.

EMAIL: lilly.yue@fda.hhs.gov

› Synthesize Real-World Data for Establishing Performance Goals in Single-Group Medical Device Clinical Studies

Chenguang Wang*, Johns Hopkins University

FDA allows sponsors to employ single-group studies for the pre-market evaluation of medical devices when the device technology is well developed and the disease of interest is well understood. Determining performance goals, i.e. the numerical target values pertaining to a effectiveness or safety endpoint, is often the critical step in such single-group studies for the FDA to make regulatory decisions. Real-world data (RWD), such as procedure or disease registries, electronic health records and insurance claim databases, provides a huge pool of subjects on which to base realistic benchmarks for determining the performance goals. In this talk, we propose a Bayesian propensity score based method for establishing performance goals using RWD in single-group medical device clinical studies.

EMAIL: cwang68@jhmi.edu

› Addressing Unmeasured Confounding in Comparative Effectiveness Research

Douglas E. Faries*, Eli Lilly and Company
Wei Shen, Eli Lilly and Company
Xiang Zhang, Eli Lilly and Company

The use of real world / observational / big data for comparative effectiveness analyses has grown in recent years. To ensure appropriate use of information arising from such comparative observational research, analyses should include a thorough and quantitative assessment of the potential impact of unmeasured confounders. The two main goals of this talk are to 1) introduce a best practice guidance for addressing unmeasured confounding; 2) demonstrate how one can incorporate information obtained external from the research study to reduce bias caused by unmeasured confounding. The best practice guidance will include a flowchart / decision tree approach to recommending analysis options given the study scenario and availability of information. The second objective focuses on the common scenario where information on confounders exists in sources external to the particular study. A Bayesian Twin Regression modeling approach will be presented that can incorporate information regarding confounders obtained from multiple external data sources and produce treatment effects adjusted for the additional information regarding key confounders.

EMAIL: shen@lilly.com

93. Recent Advances on Comparative Effectiveness Research

› Causal Inference Methods for CER: When to Use Which Methods and Why

Douglas P. Landsittel*, University of Pittsburgh
Joyce Chang, University of Pittsburgh
Andrew Topp, University of Pittsburgh
Sally C. Morton, Virginia Tech

Methodologists in comparative effectiveness research are faced with a substantial volume of literature on many different methods for causal inference and their associated properties. These methods apply across many disciplines, types of treatment strategies and types of research studies, each with potentially different objectives and causal effects of interest. Determining which methods best apply for a given scenario therefore necessitates navigating a difficult set of questions and decisions which are specific to each research study. Our Decision Tool for Causal Inference and Observational Data Analysis Methods for Comparative Effectiveness Research (DeCODe CER), which was funded through the Patient-Centered Outcomes Research Institute, attempts to provide specific guidance on which method to use when. The tool is informed by a systematic review of simulation and theoretical results, and further simulation studies assessing bias, precision, and coverage probabilities of different propensity score methods and instrumental variable approaches. We also discuss related challenges in collaboration and educational efforts, such as online resources and associated training.

EMAIL: dpl12@pitt.edu

› Developments in Multivariate Meta-Analysis

Larry V. Hedges*, Northwestern University

Statistical methods for meta-analysis typically make the assumption that estimates are independent. This is often a sensible assumption, but there are situations in which it is clearly not (e.g., when effect sizes are computed from several endpoints in a single study, or when several studies use the same sample). Dependence in other situations is less clear cut (e.g., when effect sizes come from several studies conducted in the same laboratory). Approaches that model particular dependence structures are available, but can be cumbersome to implement, particularly for less statistically sophisticated researchers. Robust variance estimation via sandwich estimators promises to be less cumbersome to implement, but depends on having a number of studies that can be unrealistically large in some applications. I will discuss some modifications to robust variance estimators that seem to provide methods with much better performance when the number of studies is small. Estimation of effective degrees of freedom in these methods may also provide a diagnostic for sample size adequacy.

EMAIL: l-hedges@northwestern.edu

Hierarchical Models for Combining N-of-1 Trials

Christopher H. Schmid*, Brown University
Youdan Wang, Brown University

N-of-1 trials are single-patient multiple-crossover studies for determining the relative effectiveness of treatments for an individual participant. A series of N-of-1 trials assessing the same scientific question may be combined to make inferences about the average efficacy of the treatment as well as to borrow strength across the series to make improved inferences about individuals. Series that include more than two treatments may enable a network model that can simultaneously estimate and compare the different treatments. Such models are complex because each trial contributes data in the form of a time series with changing treatments. The data are therefore both highly correlated and potentially contaminated by carryover. We will use data from a series of 100 N-of-1 trials in an ongoing study assessing different treatments for chronic pain to illustrate different models that may be used to represent such data.

EMAIL: christopher_schmid@brown.edu

Improving Capacity to Make Fair Comparisons of the Effectiveness and Cost-Effectiveness of Education and Social Policy Options

Rebecca A. Maynard*, University of Pennsylvania

In recent years, there has been a major push to rely on empirical evidence to guide the development and implementation of education and social policies. Concurrently, federal agencies, state governments, and private philanthropists have shifted much of their evaluation resources to studies that systematically and rigorously estimate the impacts associated with various programs, policies or practices and they have launched clearinghouses to collect, review, rate, and disseminate that evidence judged to worthy of consideration in policy. This paper discusses model cases of evidence use, as well as common (mis-)uses of evidence from even well-designed and implemented studies in the policy development process. It provides guidance for sorting, sifting, and synthesizing evidence in ways that generate meaningful comparisons across options and it illustrates the importance of both context and cost in arriving at final judgments.

EMAIL: rmaynard@gse.upenn.edu

94. Methods for Ordinal Data

Joint Model of Longitudinal Ordinal Outcome with Competing Risks Survival Analysis

Xiao Fang*, University of Texas Medical Branch
Kristopher Jennings, University of Texas Medical Branch

In longitudinal studies involves assessing ordinal disease state at a pre-decided time point, ignoring the dependence between the development and transitioning of the disease status and the potentially informative censoring event(s) may lead to bias in the estimation of covariate effects. For instance, in aging studies a common drop-out event is death. At the same time, a longitudinal study taking place over an extended period of time may lead to informative censoring event(s) that are caused by disease-related conditions. To address the problems created by these dependencies, we propose several to joint models of longitudinal status and survival: a continuous time Markov chain model which characterizes the joint dependence among events that lead to an early exit from the study, as well as a survival model in which status history is a covariate. Results from these models will be demonstrated using data from the Hispanic Established Populations for the Epidemiologic Study of the Elderly, a longitudinal survey study which aims to estimate the prevalence of and risk factors for key health conditions in older Mexican Americans.

EMAIL: xiafang@utmb.edu

Measuring Association Between Ordinal Classifications of Many Raters

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

A measure of association is a preferred summary statistic for raters' ordinal classifications, as it incorporates valuable information about the degree of disagreement present in a dataset as well as agreement. Ordered categorical scales are often employed in screening tests such as mammography to classify a patient's test result. Concerns about wide discrepancies observed between rater's classifications in common screening tests have led to large scale agreement

studies to assess levels of association and identify factors which may improve consistency between raters. Few statistical approaches exist to assess association between multiple raters using an ordinal scale in an agreement study. In this talk we describe a population-based measure of association based upon the class of generalized linear mixed models and apply the approach to recent large-scale agreement studies.

EMAIL: kerrie@bu.edu

› Sufficient Cause Interaction for Ordinal Outcomes

Jaffer Zaidi*, Harvard University
Tyler J. VanderWeele, Harvard University

VanderWeele and Robins (Biometrika 2008) derived empirical and counterfactual conditions for sufficient cause interaction between two binary exposures and a binary outcome. Sufficient cause interaction can be shown present if there exists a subpopulation for whom the outcome will occur if both exposures are present, but will not occur if either of the two exposures is absent. We extended the sufficient cause framework from binary outcomes to ordinal outcomes. Novel empirical conditions, in the form of inequality constraints on the observed data distribution, are derived for detecting sufficient cause interaction for ordinal outcomes. These inequality constraints cannot be derived by first dichotomizing our ordinal outcome, then applying the earlier inequality tests from the framework for binary outcomes. Inference based on Wald tests for the ordinal outcome inequality constraints allows us to assess the null hypothesis that there is no sufficient cause interaction. By applying Wald tests for sufficient cause interaction to the Stanford HIV drug resistance database, we discover mutations that jointly confer resistance (none, partial, full) to particular HIV drugs.

EMAIL: jzaidi@g.harvard.edu

› A Joint Marginalized Overdispersed Random Effects Model for Longitudinal Ordinal Responses

Nasim Vahabi*, Tarbiat Modares University, Iran
Anoshirvan Kazemnejad, Tarbiat Modares University, Iran
Somnath Datta, University of Florida

We consider two correlated ordinal responses which are observed longitudinally. We consider a random effects approach to account for the correlation between these two stochastic processes and make simultaneous inference; our model also accounts for temporal correlations amongst observations taken on the same subject. Another important aspect of our model is its capacity to handle data over-dispersion in order to make reliable inference. Last but not least, it is proved that certain parameters in our joint model have marginal interpretations. We investigate the statistical properties of our estimators through extensive simulation study. Finally, the methodology was applied to a real data of children failure to thrive (FTT).

EMAIL: nasim_vahabi@yahoo.com

› Mixture-Based Clustering for Ordinal Data

Daniel Fernandez*, State University of New York at Albany
Richard Arnold, Victoria University of Wellington
Shirley Pledger, Victoria University of Wellington

Many of the methods which deal with clustering in matrices of data are based on mathematical techniques such as distance-based algorithms or matrix decomposition. In general, it is not possible to use statistical inferences or select the appropriateness of a model via information criteria with these techniques because there is no underlying probability model. Additionally, the use of ordinal data is very common (e.g. Likert or pain scale). Recent research has developed a set of likelihood-based finite mixture models for a data matrix of ordinal data. This approach applies fuzzy clustering via finite mixtures to the stereotype model. Fuzzy allocation of rows, columns, and rows and columns simultaneously (biclustering) to corresponding clusters is obtained by performing the expectation-maximization (EM) algorithm and, also by Bayesian approaches (RJMCMC sampler). Examples with ordinal data sets will be shown to illustrate the application of this approach.

EMAIL: dfernandezmartinez@albany.edu

› Nonparametric Spatial Models for Clustered Ordered Periodontal Data

Dipankar Bandyopadhyay*, Virginia Commonwealth University
Antonio Canale, University of Turin

Clinical attachment level (CAL) is regarded as the most popular measure to assess periodontal disease (PD). These probed tooth-site level measures are rounded, and recorded as whole numbers (in mm) producing clustered error-prone ordinal responses. In addition, PD progression can be spatially-referenced. In this talk, we develop a Bayesian multivariate probit framework for these ordinal responses, where the cut-point parameters linking the observed ordinal CAL levels to the latent underlying disease process can be fixed in advance. The latent spatial association characterizing conditional independence under Gaussian graphs is introduced via a nonparametric Bayesian approach motivated by the probit stick-breaking process, where the components of the stick-breaking weights follow a multivariate Gaussian density with the precision matrix distributed as G-Wishart. Both simulation studies and application to a motivating PD dataset reveal the advantages of considering this flexible nonparametric ordinal framework over other alternatives.

EMAIL: dbandyop@vcu.edu

95. Epidemiologic Methods, Causal Inference and Misclassification

› Accounting for Misclassification Bias of Binary Outcomes Due to Lack of Universal Testing: A Sensitivity Analysis

Si Cheng*, University of Cincinnati
Nanhua Zhang, Cincinnati Children's Hospital Medical Center
Lilliam Ambroggio, Cincinnati Children's Hospital Medical Center
Todd Florin, Cincinnati Children's Hospital Medical Center
Maurizio Macaluso, Cincinnati Children's Hospital Medical Center

It is common that diagnostic tests for a disease are performed only in a subset of the population at higher risk, resulting in undiagnosed cases among those who do not receive the test. This poses a challenge for estimating the prevalence of the disease in the study population, and also for studying the risk factors for the disease. This problem can be considered as a missing data problem because the disease status is unknown for those who do not receive the test. We propose a Bayesian selection model which models the joint distribution of the disease outcome and whether testing was received. The sensitivity analysis allows us to assess how the association of the risk factors with the disease outcome as well as the disease prevalence changes with the sensitivity parameter. We use simulation studies to illustrate the property of the proposed method and apply the method to a pneumonia study dataset.

EMAIL: Si.Cheng@cchmc.org

› Decomposition of the Total Effect in the Presence of Multiple Mediators and Interactions

Andrea Bellavia*, Harvard School of Public Health
Linda Valeri, Harvard Medical School

Mediation analysis allows decomposing a total effect into a direct effect of the exposure on the outcome and an indirect effect operating through a number of possible hypothesized pathways. Formal definitions of direct and indirect effects in the presence of multiple mediators have been recently presented. This situation, however, may be complicated by the presence of multiple exposure-mediator and mediator-mediator interactions. In this study 1) we obtain counterfactual definitions of such interaction terms when more than one mediator is present; 2) we derive a decomposition of the total effect that unifies mediation and interaction when multiple mediators are present; and 3) we illustrate the connection between our decomposition and the 4-way decomposition of the total effect introduced in the context of a single mediator. We illustrate the properties of the proposed framework for multiple mediators and interactions, in a secondary analysis of a pragmatic trial for the treatment of schizophrenia. We employ the decomposition to investigate the interplay of side-effects and psychiatric symptoms trajectories in explaining the effect of antipsychotic on social functioning.

EMAIL: abellavi@hsph.harvard.edu

Mediation Analysis in Observational Studies Via Likelihood

Kai Wang*, University of Iowa

Mediation analysis is of great importance in social and biomedical sciences. However, some fundamental issues remain unresolved even in the basic three-factor model. Focusing on the joint likelihood of the mediator and the outcome, this paper provides a systematic treatment on important issues such as effect quantification and inference of effect size. Within this framework, a larger effect is decomposed into several meaningful parts each of which is non-negative. Such decompositions are critical in defining relative effect size. Importantly, a likelihood ratio test (LRT) is derived for testing the mediated effect. This test is closely related to the intersection-union test (IUT) and is shown to be size-α. Furthermore, a uniformly more powerful improvement over LRT and IUT is provided. Since the proposed approach is likelihood-based, both the mediator and the outcome can be fitted with generalized linear models and other models that generate likelihood. Performance of this novel approach is demonstrated using simulated data and data from an empirical study. This method is implemented in a freely available R package.

EMAIL: kai-wang@uiowa.edu

Simplifying and Contextualizing Sensitivity to Unmeasured Confounding Tipping Point Analyses

Lucy D'Agostino McGowan*, Vanderbilt University
Robert A. Greevy, Vanderbilt University

The strength of evidence provided by epidemiological and observational studies is inherently limited by the potential for unmeasured confounding. Thus, we would expect every observational study to include a quantitative sensitivity to unmeasured confounding analysis. However, we reviewed 90 recent studies with statistically significant findings, published in top tier journals, and found 41 mentioned the issue of unmeasured confounding as a limitation, but only 4 included a quantitative sensitivity analysis. Moreover, the rule of thumb that considers effects 2 or greater as robust can be misleading in being too low for studies missing an important confounder and too high for studies that extensively control for confounding. We simplify the seminal work of Rosenbaum and Rubin (1983) and Lin, Pstay, and Kronmal (1998). We focus on three key quantities: the observed bound of the confidence interval closest to the null, a plausible residual effect size for an unmeasured binary confounder, and a realistic prevalence difference for this hypothetical confounder. We offer guidelines to researchers for anchoring the tipping point analysis in the context of the study and provide examples.

EMAIL: ld.mcgowan@vanderbilt.edu

Incorrect Inference in Prevalence Trend Analysis Due to Misuse of the Odds Ratio

Randall H. Rieger*, West Chester University
Scott McClintock, West Chester University
Zhen-qiang Ma, Pennsylvania Department of Health

Because public health agencies usually monitor health outcomes over time for surveillance, program evaluation, and policy decisions, a correct health outcome trend analysis is vital. If the analysis is done incorrectly and/or results are misinterpreted, a faulty trend analysis could jeopardize key aspects of public health initiatives such as program planning, implementations, policy development and clinical decision making. It is essential then that accurate health outcome trend analysis be implemented in any data-driven decision making process. Unfortunately, there continues to be common statistical mistakes in prevalence trend analysis. In this paper, using recently published results from the Pediatric Nutrition Surveillance System, we will show the effect that an incorrect trend analysis and subsequent interpretation of results can have. We will also propose more appropriate statistical processes, such as the log-binomial model, for these situations.

EMAIL: rrieger@wcupa.edu

An Approach to Create Valid and Strong Instrumental Variables for the Purpose of Causal Inference in Large Scales

Azam Yazdani*, University of Texas Health Science Center at Houston
Akram Yazdani, Mount Sanai
Ahmad Samiei, Hasso Plattner Institute
Eric Boerwinkle, University of Texas Health Science Center at Houston

Identification of causal relationships among variables of interest provides mechanistic understanding through revealing patterns and major drivers of the system under consideration. An established approach to identify causal relationships is application of instrumental variables (IVs). However, in genome studies, application of weak and invalid IVs happens frequently which ends in unstable results. I will review different IV methods and specifically focus on the GDAG algorithm, Genome granularity Directed Acyclic Graphs. The GDAG algorithm creates valid and strong IVs by extracting comprehensive information across the genome. Therefore, the algorithm identifies robust directions, i.e. the flow of information, and achieves causal inference in large scale-observational data. In an application, information from 1,034,945 single nucleotide polymorphisms scattered across the genome was extracted to create strong and valid IVs. In total, 788 IVs were selected to identify a causal network among 122 serum metabolites. Further analyses of the metabolomic network provided insights into metabolome relationships and led to optimal inference decisions.

EMAIL: azam.yazdani@uth.tmc.edu

A Constrained Covariance Modeling Approach for Estimation of Marginal Age Trends in the Presence of Endogenous Medication Use

Andrew J. Spieker*, University of Pennsylvania

Endogenous medication use can pose challenges when seeking to estimate the association between a predictor and the natural history of a biomarker that would have occurred in the absence of treatment. In cross-sectional data, structural equation modeling can be used to correct endogeneity bias. One way of generalizing these models to estimate marginal trends in longitudinal data is to obtain a parameter estimate acting as though observations are independent, following up with a robust variance estimator. Though this approach is valid as a Z-estimation problem, efficiency can be improved by exploiting within-subject correlation. Unrestricted covariance specification demands the computation of high-dimensional integrals with no closed form expression; approximation can be computationally taxing and render model fitting prohibitively slow. I will propose a method to specify a constrained working correlation that collapses these high-dimensional integrals into a product of single integrals, resulting in simpler estimating equations. Simulations illustrate that this model is robust and can greatly improve efficiency of estimation over the working independence approach.

EMAIL: aspieker@upenn.edu

96. Bayesian Methods, Clinical Trials and Biopharmaceutical Applications

Big Data / Re-Used Data: Example of Patients' Recruitment Modeling and Feasibility of Clinical Trials

Nicolas J. Savy*, University of Toulouse III

The duration of the recruitment in clinical trials is a question of practical paramount interest. The Poisson-Gamma model, which consists in modeling the dynamic of each centre by a Poisson process with gamma-distributed rate, has shown its relevancy on many clinical trials. The parameters of the gamma distribution are usually estimated from recruitment data collected at an interim time. The problem is at the designing stage, when there is no information available to calibrate the model. Relevant scenarios of recruitment dynamics may be proposed using recruitment data from completed trials. By means of IFM 2005 and IFM 2009 (completed trials) and IFM 2014 (trial on progress during these investigations) the performances of Poisson-gamma model in that context are assessed and exhibits very good properties. In these times of Big Data, much interest is devoted to develop tools for dealing with huge databases rather than to develop methodology for re-use data of completed trials.

These investigations show that, in this context, a rational re-use of clinical data may yield to relevant results. Quality is sometime (often) better than quantity.

EMAIL: Nicolas.Savy@math.univ-toulouse.fr

› Simulation Study of Several Bayesian Approaches to Phase I Clinical Trials

Hong Wang*, University of Pittsburgh

In phase I cancer trials, it is important to effectively treat patients and minimize the chance of exposing them to sub-therapeutic and overly toxic doses. Bayesian approaches to finding the maximum tolerated dose in phase I cancer trials is discussed. The approaches relies on a realistic dose-toxicity model, allows one to include prior information. In this talk, several Bayesian methods are discussed, and compared with a simulation study.

EMAIL: how8@pitt.edu

› Bregman Divergence to Generalize Bayesian Influence Measures for Data Analysis

Matthew Weber*, Florida State University
Debajyoti Sinha, Florida State University
Dipak Dey, University of Connecticut
Luis Castro, Universidad de Concepcion

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

EMAIL: mweber022@gmail.com

› A Bayesian Bivariate Mixed Response Model with Multiplicity Adjustment for Biopharmaceutical Applications

Ross Bray*, Eli Lilly and Company
John Seaman, Baylor University
James Stamey, Baylor University

Consider a bivariate model for associated binary and continuous responses such as those in a clinical trial where both safety and efficacy are observed. We designate a marginal and conditional model that allows for the association between the responses by including the marginal response as an additional predictor of the conditional response. We use a Bayesian approach to model the bivariate regression using a hierarchical prior structure and explore various multiplicity adjustment methods within the context of the model. Of course, operating characteristics such as power and type I error rates are repeated sampling constructs, averaging, as it were, over the sample space. Bayesian methods condition on the data, averaging over the parameter space. However, in a biopharmaceutical context, regulatory requirements necessitate a thorough examination of a model's operating characteristics. To do this, we explore the so-called null space to verify consistent model performance. We also investigate the model's operating characteristics under various adjustments for multiplicity, demonstrating improved power compared to overly conservative Bonferroni adjustments.

EMAIL: bray_ross@lilly.com

› Sparse Bayesian Nonparametric Regression with Application to Health Effects of Pesticides Mixtures

Ran Wei*, North Carolina State University
Brian Reich, North Carolina State University
Jane Hoppin, North Carolina State University
Subhashis Ghosal, North Carolina State University

In many epidemiological studies that simultaneously investigate the effect of several exposure variables, the statistical challenge is to identify the subset of the exposures that affect the response significantly and to estimate the joint effects of multiple exposures. We propose a Bayesian nonparametric regression model and use continuous shrinkage priors for

variable selection and prediction. Our general approach is to decompose the exposure-response function as the sum of nonlinear main effects and two-way interaction terms, and apply the proposed computationally-advantageous continuous shrinkage Bayesian variable selection method to identify important effects. Theoretical studies show strong asymptotic estimation and variance selection properties, while numerical simulations demonstrate model performance under practical scenarios. The method is applied on neurobehavioral data from Agricultural Health Study that investigates the associations between pesticide use and neurobehavioral outcomes in farmers.

EMAIL: rwei@ncsu.edu

› A Bayesian Design for Two-Arm Randomized Biosimilar Clinical Trials

Haitao Pan*, University of Texas MD.Anderson Cancer Center
Ying Yuan, University of Texas MD.Anderson Cancer Center

A biosimilar refers to a follow-on biologic intended to be approved for marketing based on biosimilarity to an existing patented biological product. To develop a biosimilar product, it is essential to demonstrate biosimilarity between the follow-on biologic and the reference product, typically through two-arm randomization trials. We propose a Bayesian adaptive design for trials to evaluate biosimilar products. To take advantage of the abundant historical data on the efficacy of the reference product that is typically available at the time a biosimilar product is developed, we propose the calibrated power prior, which allows our design to adaptively borrow information from the historical data according to the congruence between the historical data and the new data collected from the current trial. We propose a new measure, the Bayesian biosimilarity index. We used the proposed methods in a group sequential fashion based on the accumulating interim data and stop the trial early once there is enough information to conclude or reject the similarity. Extensive simulation studies show that the proposed design has higher power than traditional designs.

EMAIL: davidhaitaopan@gmail.com

› A Group Sequential Test for Treatment Effect Based on the Fine-Gray Model

Michael J. Martens*#, Medical College of Wisconsin
Brent R. Logan, Medical College of Wisconsin

Competing risks endpoints arise when patients can fail therapy from several causes. Analyzing these outcomes allows one to assess the direct benefit of treatment on a primary cause of failure in a clinical trial setting. Regression models can be used in clinical trials to adjust for residual imbalances in patient characteristics, improving the power to detect a treatment effect. But, none of the currently available group sequential tests for competing risks adjusts for covariates. We propose a group sequential test for treatment effect based on the Fine-Gray model that permits covariate adjustment. Its test statistics have an asymptotic distribution with an independent increments structure, allowing usage of standard techniques such as error spending functions to meet type I error rate and power specifications. We demonstrate the test in a reanalysis of a phase III trial that evaluated an experimental treatment for the prevention of adverse outcomes following blood and marrow transplant. Moreover, a comprehensive simulation study demonstrates that the proposed method preserves the type I error rate and power at their nominal levels in the presence of influential covariates.

EMAIL: mmartens@mcw.edu

97. Estimating Quation Approaches

› Secondary Outcome Analysis for Data from an Outcome-Dependent Sampling Design

Yinghao Pan*, University of North Carolina, Chapel Hill
Jianwen Cai, University of North Carolina, Chapel Hill
Matthew P. Longnecker, National Institute of Environmental Health Sciences, National Institutes of Health
Haibo Zhou, University of North Carolina, Chapel Hill

For a study with continuous primary outcome, Zhou et al. (2002) considers an outcome-dependent sampling (ODS) scheme, where the expensive exposure is measured on a simple random sample and supplemental samples selected based on the primary outcome. With tremendous cost invested in collecting the

primary exposure information, investigators often want to utilize the available data to study the relationship between a secondary outcome and the obtained exposure variable. This is referred as secondary analysis. Secondary analysis in ODS designs can be tricky, as the ODS sample is not a simple random sample. In this article, we use the inverse probability weighted (IPW) and augmented inverse probability weighted (AIPW) estimating equations to analyze the secondary outcome for data obtained from the ODS design. We do not make any parametric assumptions on the primary and secondary outcome, and only specify the form of the regression mean, thus avoids the possibility of model misspecification. Our proposed estimator is shown to be consistent and efficient. Data from the Collaborative Perinatal Project (CPP) is analyzed to illustrate our method.

EMAIL: yypan@live.unc.edu

› Two-Sample Tests for Quantile Residual Life Time

Yimeng Liu*, University of Pittsburgh
Abdus S. Wahed, University of Pittsburgh
Gong Tang, University of Pittsburgh

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

EMAIL: cn.yimeng@gmail.com

› Stagewise Generalized Estimating Equations with Grouped Variables

Gregory Vaughan*, University of Connecticut
Robert Aseltine, University of Connecticut Health Center
Kun Chen, University of Connecticut
Jun Yan, University of Connecticut

Forward stagewise estimation is a slow-brewing approach for model building that is attractive in dealing with complex data structures for both its computational efficiency and its connections with regularization. Under the framework of generalized estimating equations, we study general stagewise estimation approaches that can handle clustered data and non-Gaussian models with a variable grouping structure. In practice the group structure is often not ideal; the key is to conduct both group and within-group variable selection. We propose two approaches to address the challenge. The first is a bi-level stagewise estimating equations (BiSEE) approach, which is shown to correspond to the Sparse group lasso penalized regression. The second is a hierarchical stagewise estimating equations (HiSEE) approach to handle more general hierarchical grouping structures. Simulation studies show that BiSEE and HiSEE are competitive compared to existing approaches. We apply the proposed approaches to study the association between the suicide-related hospitalization rates of the 15-19 age group and the characteristics of the school districts in the State of Connecticut.

EMAIL: gregory.vaughan@uconn.edu

› Variable Selection for Correlated Bivariate Mixed Outcomes Using Penalized Generalized Estimating Equations

Ved Deshpande*, University of Connecticut
Dipak K. Dey, University of Connecticut
Elizabeth D. Schifano, University of Connecticut

We propose a penalized generalized estimating equations framework to jointly model correlated bivariate binary and continuous outcomes involving multiple predictor variables. We use sparsity-inducing penalty functions to simultaneously estimate the regression effects and perform variable selection

on the predictors, and use cross-validation to select the tuning parameters. We further propose a method for tuning parameter selection that can control a desired false discovery rate. Using simulation studies, we demonstrate that the proposed joint modeling approach performs better in terms of accuracy and variable selection than separate penalized regressions for the binary and continuous data. We demonstrate the application of the method on a medical expenditure data set.

EMAIL: ved.deshpande@uconn.edu

› Kernel-Based Causal Inference

Yuying Xie*, University of Waterloo
Yeying Zhu, University of Waterloo
Cecilia A. Cotton, University of Waterloo

It is shown that kernel distance is one of the best bias indicator in estimating the causal effect compared to other balance measures, such as absolute standardized mean difference (ASMD) and KS statistic. An important goal in estimating the causal effect is to achieve balance in the covariates. We propose both parametric and nonparametric causal effect estimators using kernel distance. The estimating equations are solved by generalized method of moments or empirical likelihood. Simulation studies are conducted across different scenarios varying in the degree of nonlinearity. The simulation study shows that the proposed approaches produce smaller mean squared errors in estimating causal treatment effects than many existing approaches including the well-known CBPS approach.

EMAIL: y63xie@uwaterloo.ca

› Joint Modeling of Longitudinal and Survival Outcomes Using Generalized Estimating Equations

Mengjie Zheng*, Indiana University School of Medicine
Sujuan Gao, Indiana University School of Medicine

The joint modeling framework for longitudinal and time-to-event data has been introduced to study the association between time-dependent covariates and the risk of an event. Existing estimation methods include the two-stage approach, Bayesian and maximum likelihood estimation methods (MLEs). The two-stage method is computationally straightforward but introduces biases, while the Bayesian and MLEs rely on the joint likelihood of longitudinal and survival processes and can be computationally intensive. In this paper, we propose a joint generalized estimating equation framework using an inverse intensity weighting approach to correct biases from the naive two-stage method. The proposed method can handle longitudinal outcomes from the exponential family of distributions and is computationally efficient to carry out. The performance of the proposed method is assessed through simulation studies. The proposed method is applied to a data from a longitudinal cohort to determine the association of longitudinal Low-density lipoprotein (LDL), High-density lipoprotein (HDL) measures and the risk of coronary artery disease (CAD).

EMAIL: zhengmen@iupui.edu

98. Graphical Models and Statistical Graphics

› Two New Residuals in Survival Analysis with Full Likelihood

Susan Halabi*, Duke University
Sandipan Dutta, Duke University

Residuals in the proportional hazards (PH) model are useful in detecting outliers or influential points, in testing the proportional hazards assumption and exploring functional form. Assuming proportional hazards and non-informative censoring, the full likelihood approach is used to obtain score and deviance residuals. The first residual is based on the ideas used in obtaining the score-type residuals in partial likelihood approach. The second type of residual is based on the concept of the deviance residuals. We conduct simulations and compare the performance of the full likelihood residuals with other common residuals that are based on the partial likelihood approach. In addition, the graphical approaches are used to illustrate the applications of these residuals using some real life examples.

EMAIL: susan.halabi@duke.edu

Functional Exponential Random Graph Models for Dynamic Networks with Temporal Heterogeneity

Jihui Lee*#, Columbia University
Gen Li, Columbia University
James D. Wilson, University of San Francisco

Network topology evolves through time. It is important to understand fluctuations in network structure to provide meaningful interpretation of dynamic networks. We propose a method, functional exponential random graph model (FERGM), which assumes smoothly transitioning networks. From a series of unweighted networks, the FERGM captures temporal heterogeneity of network topology. We express the heterogeneity with basis splines. Basis expansion simplifies the estimation, which leads to computational efficiency. Furthermore, it enables FERGM to accommodate irregular time points by borrowing strength from neighboring time points. We derive that calculating maximum pseudo-likelihood estimator (MPLE) is equivalent to fitting a penalized logistic regression. We provide the method of iteratively reweighted least squares (IRLS) with smoothing penalty. We conduct a simulation study to evaluate the performance of FERGM in terms of capturing how network topology fluctuates through time. The FERGM is also applied to resting state fMRI data. In both simulation and application, we observe that the FERGM efficiently captures smooth temporal heterogeneity of dynamic networks.

EMAIL: jl4201@cumc.columbia.edu

Visualization and Testing of Residual Correlation Patterns Using Spatial Statistics Approaches

Masha Kocherginsky*, Northwestern University
Raeed Chowdhury, Northwestern University
Lee Miller, Northwestern University

Firing rates (Y) of neurons in the primary somatosensory cortex of a monkey's brain during reaching can be modeled as a nonlinear function of the direction of limb movement (X1), direction of endpoint interaction forces (X2), and possibly their interaction. Graphical exploration of the (X1, X2) space using two-way heatmaps can be misleading if X1 and X2 are measured on an irregular grid, as interpolation in areas with sparse (X1, X2) data can appear as hotspots. We propose to use approaches from spatial statistics to examine residual patterns and test their independence. First, we use Voronoi diagrams to partition the (X1, X2) space into regions of points that are closer to (X1, X2) than any other observed pair, and shade the regions using a color palette with intensity proportional to the magnitude of the residual at (X1, X2). We then use Moran's I, a global index of spatial autocorrelation, to quantify the strength of spatial correlation and test for its statistical significance. The novelty of this approach is in the representation of the combined force-velocity data as a spatial problem, and applying spatial statistics methods to assess residual structure.

EMAIL: mkocherg@northwestern.edu

A Convex Framework for High-Dimensional Sparse Cholesky Based Covariance Estimation in Gaussian DAG Models

Kshitij Khare, University of Florida
Sang Oh, University of California, Santa Barbara
Syed Hafizur Rahman*, University of Florida
Bala Rajaratnam, Stanford University

Covariance estimation for high-dimensional datasets is a fundamental problem in modern day statistics with numerous applications. In these high dimensional datasets, the number of variables p is typically larger than the sample size n. A popular way of tackling this challenge is to induce sparsity in the covariance matrix, its inverse or a relevant transformation. In particular, methods inducing sparsity in the Cholesky parameter of the inverse covariance matrix can be useful as they are guaranteed to give a positive definite estimate of the covariance matrix. Also, the estimated sparsity pattern corresponds to a Directed Acyclic Graph (DAG) model for Gaussian data. In this paper, we propose a new penalized likelihood method for sparse estimation of the inverse covariance Cholesky parameter that aims to overcome some of the shortcomings of current methods, but retains their respective strengths. We obtain a jointly convex formulation for our objective function, which leads to convergence guarantees, even when $p > n$. We establish high-dimensional estimation and graph selection consistency, and also demonstrate finite sample performance on simulated/real data.

EMAIL: shr264@ufl.edu

Joint Estimation of Panel VAR Models Sharing Common Structure

Andrey V. Skripnikov*, University of Florida
George Michailidis, University of Florida

In a number of applications, one has access to mutlivariate panel time series data on K units that share common structure. In this work, we discuss the problem of their joint estimation so as to increase statistical efficiency of the parameters of the common structure leveraging a group lasso penalty. Further, we examine the case where the K units can be divided into groups and one is interested in testing for differences in the parameters across groups. The results are illustrated on synthetic and real data sets.

EMAIL: usdandres@ufl.edu

Inference on Large Scale Bayesian Network

Suwa Xu*, University of Florida

We propose partial p-value algorithm, a simple and computationally-efficient algorithm for learning large scale Bayesian networks. By specifying that each node-conditional distribution is a member of exponential family, the problem of estimating the structure of Bayesian network amounts to estimating the regression coefficients of generalized linear models. The recovery of the network structure involves two screening steps, one is obtained by thresholding the marginal regression coefficients and the other is obtained by thresholding the reduced size regression coefficients. A multiple hypothesis test-based method is used in these two screening steps. Furthermore, we add edge directions based on v-structure and stochastic simulated annealing. Theoretically, we establish the consistency of the proposed methods under mild conditions, showing that with high probability, the edge structure of our graphical models can be recovered exactly. We illustrate the performance of our method in two simulation studies and on a real genomic dataset, and show that it performs quite competitively in practice.

EMAIL: suwaxu@ufl.edu

99. Data-Adaptive Modeling in Causal Inference

A Self-Selecting Procedure for the Optimal Discretization of the Timeline for Longitudinal Causal Inference Methods with Electronic Health Data

Steve Ferreira Guerra, Universite de Montreal
Mireille E. Schnitzer*, Universite de Montreal
Amelie Forget, Universite de Montreal
Lucie Blais, Universite de Montreal

In longitudinal observational studies, marginal structural models (MSMs) can be used to account for time-dependent confounding. Estimation approaches, such as Longitudinal Targeted Maximum Likelihood Estimation (L-TMLE), often require a finite number of time points where treatment exposure and covariates are measured. In contrast, electronic health data (EHD) is produced by mechanisms that collect health system user information in real-time. In common practice, the application of longitudinal causal estimators is preceded by an arbitrary discretization of a continuous timeline. In this talk, we describe the causal inference problem when the operating data is defined as a coarsening (discretization) of the observed data. We then consider the impacts of discretization on the interpretation of a MSM parameter and on estimation. We then propose a novel selection procedure that uses cross-validation on an L-TMLE loss function to select an "optimal" discretization for estimation of a data-adaptive MSM parameter of interest. This approach is applied to a study using EHD to evaluate the impact of similarly indicated asthma medications during pregnancy on pregnancy duration.

EMAIL: mireille.schnitzer@umontreal.ca

Improved Doubly Robust Estimation in Longitudinal Missing Data and Causal Inference Models

Alexander R. Luedtke*, Fred Hutchinson Cancer Research Center
Mark J. van der Laan, University of California, Berkeley
Marco Carone, University of Washington

Bang and Robins, and subsequently van der Laan and Gruber, provided sequential estimators for the G-formula for

the longitudinal counterfactual mean outcome. While both approaches are double robust in that they are consistent if either the outcome regression or the treatment mechanism is consistently estimated at each time point, this statement turns out to be misleading for sequential regression estimators. Though consistent estimation of the treatment mechanism at one time point does not rely on consistent estimation of the treatment mechanism other time points, consistently estimating an outcome regression for these sequential estimators requires consistently estimating the outcome regression at each subsequent time point. We introduce an outcome regression estimator that is more robust than earlier proposals: rather than requiring consistent estimation of all subsequent outcome regressions, it requires that, at each subsequent time point, either the outcome regression or the treatment mechanism is consistently estimated. Developing this procedure involves a novel translation of ideas used in targeted minimum loss-based estimation to the infinite-dimensional setting.

EMAIL: aluedtke@fredhutch.org

› Optimal Nonparametric Estimation and Testing of Treatment Effect Heterogeneity

Edward H. Kennedy*, Carnegie Mellon University

The variance of the individual treatment effect is a fundamental quantity in causal inference and is important in a wide array of settings. It may be of interest in its own right, as it represents an essential component of the intervention/population under study; it also plays a crucial role in design-based inference for sample average treatment effects, and can help determine optimal treatment regimes. Unfortunately, the treatment effect variance is not identified even in a randomized trial, since it depends on the joint distribution of potential outcomes; however we can construct bounds. Previous work has only considered settings without any covariates, but high-dimensional covariate information is often available and can provide much tighter and more informative bounds. In this work we develop nonparametric methods for estimating covariate-adjusted bounds on the treatment effect variance, which can incorporate machine learning while still providing valid inference. We also present higher-order theory, which is crucial for testing constant effect hypotheses and for minimax estimation in settings where root-n convergence rates cannot be attained.

EMAIL: edwardh.kennedy@gmail.com

100. Advances and New Directions in the Use of Bayesian Methods In Pharmaceutical Development

› The Value of Bayesian Approaches in the Regulation of Medical Products

Telba Z. Irony*, U.S. Food and Drug Administration

Regulatory decision making for approval of medical products involves benefit-risk assessments of all available information about the benefits and risks of a treatment. A comprehensive benefit-risk determination should also take into account other factors reflecting the context in which these assessments are made. These factors include, among others, the uncertainty about the benefits and risks, the availability of alternative treatments, and the option of taking preliminary action and deferring a final benefit-risk determination to a later time, once more information is acquired. While the main challenge of benefit-risk determinations for medical product approval is to combine information with values, the essence of the Bayesian approach is to collect and update information, and merge it with values to make rational decisions. In this talk I will argue that the Bayesian approach provides an ideal framework for benefit-risk assessments for medical products and provides examples where Bayesian approaches are particularly advantageous in the regulatory setting.

EMAIL: telba.irony@fda.hhs.gov

› Efficient Decision-Making for Clinical Drug Development

Stephen J. Ruberg*, Eli Lilly and Company
Margaret Gamalo, Eli Lilly and Company
Karen Price, Eli Lilly and Company
Zongjun Zhang, Eli Lilly and Company

When clinical drug development is cast in a frequentist paradigm, early phase trials are hypothesis generating trials. When

positive, these are followed by confirmatory trials generally using a significance level of 0.05 for hypothesis testing This process is done to satisfy the well-known convention for "substantial evidence." This does not allow the direct use of all data in the decision-making process as to whether a drug should be approved or not. For example, an extensive early phase program with very credible results is not combined with confirmatory trial data to interpret the significance of a treatment's efficacy. When the cost and duration of drug development is increasing, the scientific community should revisit the current standard for substantial evidence and allow Bayesian approaches to be formally incorporated into the drug development and approval process. This presentation will explore how that might be implemented practically without loss of scientific rigor or relaxing the high standard needed for treatments to be approved for marketing as well as posit some efficiency gains from Bayesian approaches.

EMAIL: sruberg@lilly.com

› A Pharmaceutical Industry Perspective on the Value of Bayesian Methods

David I. Ohlssen*, Novartis Pharmaceuticals Corporation

In the first part of the talk, we will briefly review the status of Bayesian methods in drug development. Emphasis will be placed on the education efforts required to move Bayes into practice, together with examples illustrating a number of success stories. The second part of the talk will look at challenges to the broader use of Bayesian methods, particularly in the regulatory setting. We will provide a balanced discussion on the definition of substantial evidence used for regulatory approval, the use of simulated type one error control in designs with Bayesian decision rules, Bayesian decision making to evaluate benefit risk and the use of Bayesian thinking without formal Bayesian inference.

EMAIL: david.ohlssen@novartis.com

› Bayesian Analyses in Confirmatory Trials: What Has Been Done and What Can Be Done?

Scott M. Berry*, Berry Consultants

There is a general impression throughout the drug development community that Bayesian analyses are not being used in pivotal trials in a regulatory environment. In this talk I will describe the current state of Bayesian methods being used in pivotal trials and also touch on what the future realistically holds for Bayesian methods being used. Multiple examples will be presented of actual pivotal trials and future problems will be posed where Bayesian methods offer a real solution.

EMAIL: scott@berryconsultants.com

101. Design and Analysis of Cancer Immunotherapy Trials

› Designing Therapeutic Cancer Vaccine Trials with Random Delayed Treatment Effect

Zhenzhen Xu*, U.S. Food and Drug Administration
Boguang Zhen, U.S. Food and Drug Administration
Yongseok Park, University of Pittsburgh
Bin Zhu, National Cancer Institute, National Institutes of Health

Arming the immune system against cancer has emerged as a powerful tool in oncology during recent years. Instead of poisoning a tumor or destroying it with radiation, therapeutic cancer vaccine, a type of cancer immunotherapy, unleashes the immune system to combat cancer. This indirect mechanism-of-action of vaccines poses the possibility of a delayed onset of clinical effect. This results in a delayed separation of survival curves between the experimental and control groups in therapeutic cancer vaccine trials with time-to-event endpoints and violates the proportional hazard assumption. As a result, the conventional study design based on the regular log-rank test ignoring the delayed effect would lead to a loss of power. In addition, the delayed duration seems to vary from subject to subject or from study to study, and thus it could be considered as random. In this talk, we propose innovative approaches for sample size and power calculation to properly and efficiently account for the delayed effect with random duration into the design and analysis of such a trial and evaluate the proposed approaches both analytically and empirically.

EMAIL: zhenzhen.xu@fda.hhs.gov

› Design and Analysis of Clinical Trials in the Presence of Delayed Treatment Effect

Tony Sit, The Chinese University of Hong Kong
Mengling Liu*, New York University School of Medicine
Michael Shnaidman, Pfizer Inc.
Zhiliang Ying, Columbia University

In clinical trials with survival endpoint, it is common to observe an overlap between two Kaplan–Meier curves of treatment and control groups during the early stage of the trials, indicating a potential delayed treatment effect. Formulas have been derived for the asymptotic power of the log-rank test in the presence of delayed treatment effect and its accompanying sample size calculation. We will present a reformulated the alternative hypothesis with the delayed treatment effect in a rescaled time domain, which can yield a simplified sample size formula for the log-rank test in this context. We further propose an intersection-union test to examine the efficacy of treatment with delayed effect and show it to be more powerful than the log-rank test. Simulation studies are conducted to demonstrate the proposed methods.

EMAIL: mengling.liu@nyumc.org

› Interim Analysis for Time to Event Trial with Delayed Treatment Effect

Satrajit Roychoudhury*, Novartis Pharmaceuticals Corporation

Innovative research in recent years has led to the discovery of targeted immunotherapies such as monoclonal antibodies, T cell infusion, and cancer vaccines. Immunotherapies, on the other hand, stimulate the patient's own immune system to fight against cancers by targeting antigens expressed on cancer cells. Several immunotherapies have been demonstrated to show long term survival and delayed clinical benefit. This poses challenge to the proportionality hazard assumption that is the key for most of the commonly used study designs (e.g. group sequential) and analysis methods used for time to event endpoint. If the treatments exhibit delayed clinical benefit, implementation of efficacy interim analysis may have smaller stopping probability for a positive outcome whereas futility interim analysis could increase the chance of terminating the study early and erroneously discarding an active agent. We investigated different methodologies for interim analysis. Benefits and limitations of different approaches are explored using simulation studies.

EMAIL: satrajit.roychoudhury@novartis.com

102. New Developments of Small Area Estimation Models

› Global-Local Priors for Small Area Estimation

Xueying Tang, University of Florida
Malay Ghosh*, University of Florida

The paper introduces global-local shrinkage priors for random effects in small area estimation. These priors had earlier success in Bayesian multiple testing and Bayesian variable selection since they can capture potential sparsity, which in the present context means lack of significant contributions by most of the random effects. In addition, these priors can quantify and assess disparity in the area level random effects. The basic idea is to employ two levels of parameters to express random effect variances. One is the global shrinkage parameter, which is common for all the random effects and introduces an overall shrinkage effect. The other is the local shrinkage parameter, which acts at an individual level and prevents overshrinkage. We show via simulations and data analysis that the global-local priors work as well or even better than some of the other priors proposed earlier.

EMAIL: ghoshm@stat.ufl.edu

› A Simple Adaptation of Variable Selection Software for Regression Models to Select Variables in Nested Error Regression Models

Yan Li*, University of Maryland, College Park
Partha Lahiri, University of Maryland, College Park

Data users often apply standard regression model selection criterion to select variables in nested error regression models, which are widely used in small area estimation. We demonstrate through a Monte Carlo simulation study that this practice may lead to selection of a non-optimal or incorrect model. To assist data users who wish to use standard regres-

sion software, we propose a transformation of the data so that transformed data follow a regression model. Thus, standard variable selection software available for the regression model can be applied to the transformed data. We illustrate our methodology using survey and satellite data for corn and soybeans in 12 Iowa counties.

EMAIL: yli6@umd.edu

› Spatio-Temporal Change of Support with Application to American Community Survey Multi-Year Period Estimates

Scott H. Holan*, University of Missouri
Jonathan R. Bradley, Florida State University
Christopher K. Wikle, University of Missouri

Motivated by the American Community Survey (ACS), we present Bayesian methodology to perform spatio-temporal change of support (COS) for survey data with Gaussian sampling errors. The ACS has published 1-year, 3-year, and 5-year period estimates, and margins of errors, for demographic and socio-economic variables recorded over predefined geographies. The spatio-temporal COS methodology considered here provides users a way to estimate ACS variables on customized geographies and time periods while accounting for sampling errors. Additionally, 3-year ACS period estimates will be discontinued, and this methodology can provide predictions of ACS variables for 3-year periods given the available period estimates. The methodology is based on a spatio-temporal mixed-effects model with a low-dimensional spatio-temporal basis function representation, which provides multi-resolution estimates through basis function aggregation in space and time. This methodology includes a novel parameterization that uses a target dynamical process. The effectiveness of our approach is demonstrated through two applications using public-use ACS estimates and is shown to produce good predictions.

EMAIL: holans@missouri.edu

› Bayesian SAE Modeling with Benchmarking and Spatial Smoothing for Estimating Health Policy Impact

Hanjun Yu, Peking University
Zhaonan Li, Peking University
Xinyi Xu*, The Ohio State University
Bo Lu, The Ohio State University

Small area estimation (SAE) concerns with how to reliably estimate population quantities of interest when some areas or domains have very limited samples. This is an important issue in large population surveys, because the geographical areas or groups with only small samples are often of great interest to researchers and policy makers. Classic approaches usually provide accurate estimators at state level or large geographical region level, but they fail to provide reliable estimators for many rural counties where the samples are sparse. These measures are crucial to evaluating the health policy impact locally. We propose a Bayesian hierarchical model with constraints on the parameter space, which is shown to outperform the traditional benchmarking method. We also incorporate different spatial smoothing strategies to take advantage of neighborhood information. Causal interpretation for evaluating health policy is discussed. We apply our model to 2015 Ohio Medicaid Assessment Survey (OMAS) data to investigate the county level impact of Affordable Care Act (ACA) on insurance and health outcomes.

EMAIL: xinyi@stat.osu.edu

103. Harnessing Big Data for Phenotype Estimation and Analysis

› Statistical Solutions for Risk Prediction with Imperfect Phenotypes Derived from Electronic Medical Records

Jennifer A. Sinnott*, The Ohio State University

Electronic medical records provide a powerful new data resource that can provide much larger, more diverse samples for common tasks such as genetic association testing and risk prediction model building. However, algorithms for estimating certain phenotypes, especially those that are complex and/or difficult to diagnose, produce outcomes and covariates

subject to measurement error. Much work is needed to determine best practices for implementing and analyzing such data. To this end, we recently proposed a method for analyzing case-control studies when disease status is estimated by a phenotyping algorithm; our method improves power and eliminates bias when compared to the standard approach of dichotomizing the algorithm prediction and analyzing the data as though case-control status were known perfectly. The method relies on knowing certain parameters describing the algorithm, but in practice these may not be known, and we discuss extending the method to be more robust to misspecification of these parameters. We also discuss statistical methods when both the outcome and some predictors are derived from imperfect phenotyping algorithms.

EMAIL: jsinnott@stat.osu.edu

› Predicting Multiple Outcomes with Semiparametric Canonical Correlation Analysis

Denis Agniel*, RAND Corporation
Tianxi Cai, Harvard University

When studying complex disorders in electronic health records and other passive data sources, a single, well-defined measurement may not be available to accurately describe the scientific question. In these cases, multiple, possibly diverse outcomes may be ascertained which describe aspects of the phenotype researchers wish to measure. The goal of this work is to build prediction models in this challenging setting. We seek specifically to identify risk scores as linear combinations of a set of predictors that are highly predictive of a set of such diverse outcomes. Further complicating matters, the outcomes may be measured on completely different scales, i.e. some outcomes may be binary, others continuous, and still others censored in some way. We use marginal parametric and semiparametric transformation models to put all outcomes on the same scale together with the canonical correlation analysis framework to build risk scores for prediction. We demonstrate the asymptotic properties of the procedure and discuss its advantages over existing methods, such as vector generalized linear models.

EMAIL: denis.agniel@mail.harvard.edu

› Tracking the Effects of Healthcare Innovations Through the Use of Electronic Health Records

Joseph Lucas*, Duke University

Changes in incentives and business models are leading health systems to rapidly innovate many aspects of care. However, the effects of those innovations are often not being tracked. In scientific terms, this is like running an experiment without paying any attention to the outcome. Changes in healthcare delivery should be treated as a full-fledged experiment; a hypothesis should be generated, tools for measuring the outcome should be developed and implementation should be conducted in a way to allow results to be measured. In this talk we will discuss some tools for ongoing tracking of healthcare innovations. In contrast to a more typical approach to trials involving fixed time intervals and sample sizes, our approach uses a Bayesian framework to continuously update our understanding of the effects of the innovation. This allows presentation of results in a context that is useful for supporting decision making by health system administrators.

EMAIL: joseph.lucas@duke.edu

› Inferring Mobility Traces from Smartphone GPS Data via Digital Phenotyping

Jukka-Pekka Onnela*, Harvard School of Public Health
Ian Barnett, Harvard School of Public Health

Large-scale phenotyping is a natural complement to genome sequencing. Of the different phenotype classes, behavior has traditionally presented special challenges for phenomics due to its temporal nature. Smartphones can contribute to this phenotyping challenge via objective measurement. We have previously defined "digital phenotyping" as the moment-by-moment quantification of the individual-level phenotype in situ using data from personal digital devices, and our smartphone-based digital phenotyping platform enables us to collect different types of active and passive data from subjects. This approach opens up some intriguing scientific opportunities, but it also presents several interesting statistical challenges. I will discuss one of them, the problem of inferring mobility measures from GPS traces with substantial missingness. Our approach to imputation, based on resam-

pling of observed mobility trajectories, appears to perform well on empirical data, and our analytical results yield further insight into its properties. Finally, I will present a longitudinal case study that uses GPS-based mobility measures in the modeling of patient outcomes.

EMAIL: onnela@hsph.harvard.edu

104. New Development in Causal Inference

Usefulness of Parental Genotypes in Mendelian Randomization

Rui Feng*, University of Pennsylvania
Wei Yang, University of Pennsylvania

Instrumental variables (IV) have been used for estimating the causal effect of an exposure variable on an outcome in observational studies. The goal of introducing IV is to remove potential selection bias or unobserved confounder effects. With genotypes as IV, the so-called Mendelian Randomization (MR) framework has recently been applied to genetic studies to understand the true effect of a biomarker on an outcome. A valid inference requires that IV should have a relatively strong association with the biomarker, but have no direct effect on the outcome and is independent with unmeasured confounders or selection probability. Either confounders or sample selection, though not directly caused by children's genes, can be correlated with them through parental genetic influences. Because parental genotypes are often available, we proposed an MR method that stratifies on parental genotypes to construct unbiased causal inference of children's biomarkers on their outcomes. The properties of our method are demonstrated through simulations and new design advice is provided.

EMAIL: ruifeng@upenn.edu

Causal Comparison of Treatment Policies using Electronic Health Records Data

Joseph W. Hogan*, Brown University
Hana Lee, Brown University
Becky Genberg, Brown University
Ann Mwangi, Moi University School of Medicine
Paula Braitstein, University of Toronto

The HIV care cascade is a framework that identifies key stages of HIV care and treatment: case identification, linkage to care, initiation of antiviral treatment, and viral suppression. UNAIDS has set targets for healthcare systems in terms of percentage of individuals reaching each stage. Cascade-based evaluations of specific policies have primarily relied on microsimulation from mathematical models. Growing availability of longitudinal electronic health records (EHR) on HIV-infected individuals presents an opportunity for more unified approaches using statistical models. We describe a framework for causal inference about treatment policies using a structural state space model of the care cascade. Causal comparisons are defined in terms of state membership probabilities, and inferred using G estimation methods. We analyze data from over 50,000 individuals in an HIV care program in Kenya, yielding causal comparisons of "test-and-treat" versus dynamic regimes based on evolving CD4 counts. The findings have important implications for treatment recommendations in resource-constrained settings.

EMAIL: jhogan@stat.brown.edu

Infererring Causal Effects with Invalid Instruments in Mendelian Randomization

Hyunseung Kang*, University of Wisconsin, Madison

Recently in a subfield of genetic epidemiology known as Mendelian randomization (MR), instrumental variables (IV) techniques have been used to estimate the causal effect of an exposure on an outcome using genetic markers as instruments. These IV techniques require (i) instruments that are strongly associated with the exposure and (ii) a complete knowledge about all the instruments' validity; a valid instrument must not have a direct effect on the outcome and not be related to unmeasured confounders. Often, this is impractical

in many MR studies where genetic instruments are weak and complete knowledge about genetic instruments' validity is absent. The talk discusses two procedures that provide honest inference of the causal effect in MR studies. The first inference procedure, called two-stage hard thresholding (TSHT), is robust to invalid IVs and achieves optimal (or near-optimal) performance. The second inference procedure is robust to both invalid and weak IVs and has uniformity guarantees. Both procedures can also handle summary-level MR data.

EMAIL: hkang.email@gmail.com

› Mediation Analysis with Latent Class Mediators

Haiqun Lin*, Yale University
Kyaw Sint, Yale University
Robert Rosenheck, Yale University

We propose latent class mediators to study the effect of intervention on outcome that is exerted through multiple mediators. We postulate that the multivariate mediators can be summarized into several latent classes with each class representing a different joint distribution of these mediators. The use of the latent class mediators enable us to avoid the modeling the interaction effects between exposure and mediators and between mediators on the outcome. We use the causal mediation framework to derive the mediation effect by the multiple mediators and direct effect of the exposure. Our method is applied to the analysis of the mediation effect of the longitudinal adherence to the four intervention components in the two-year clustered randomized clinical trial: the Recovery after an Initial Schizophrenia Episode (RAISE) Early Treatment Program (ETP). The four intervention components include personalized medication management, family psychoeducation; individual resilience therapy; and supported education and employment. We assess whether the improvement in functional outcome of quality of life is mediated through the adherence to the four intervention components.

EMAIL: haiqun.lin@yale.edu

105. ICSA Invited Session: Statistics Into Biosciences

› Latent Variable Poisson Models for Assessing Regularity of Circadian Patterns Over Time

Paul S. Albert*, National Cancer Institute, National Institutes of Health

Many researchers in biology and medicine have focused on trying to understand biological rythms and their potential for impact on disease. A common biological rhythm is circadian where the cycle repeats itself every 24 hours. However, a disturbance of the circadian pattern may be indicative of future disease. In this paper, we develop new statistical methodology for assessing the degree of disturbance or irregularity in a circadian pattern for count sequences. We develop a latent variable Poission modeling approach with both circadian and stochastic short-term trend that allow for individual variation in the proportion of these components.We propose novel Bayesian estimation approaches. We illustrate the approach with intensively collected longitudinal data from a study on adolescents.

EMAIL: albertp@mail.nih.gov

› Detection of Differentially Methylated Regions Using BS-seq Data

Shili Lin*, The Ohio State University

DNA methylation is an epigenetic change occurring in genomic CpG sequences that contribute to the regulation of gene transcription both in normal and malignant cells. Aided by fast parallel sequencing technology in recent years, a number of genome-wide platforms have been developed to provide high throughput DNA methylation data. The most well-known platforms are bisulfite conversion based, collectively known as BS-seq. Several sophisticated statistical solutions have been developed to analyze massive amounts of data produced from BS-seq. Most of the methods are proposed for detecting differentially methylated loci (DML), although differentially methylated regions (DMR) are often of more relevance biologically. The two most prominent features in BS-seq data are between-sample variability in methylation proportions and correlation in methylation signals in

neighboring CpG sites. In this talk, I will discuss a number of methods that take sample variability into account when detecting DMLs. I will also discuss a Bayesian smoothing Curve (BCurve) method to detect DMRs using credible bands. In particular, I will highlight the feature of BCurve that takes correlation into consideration.

EMAIL: shili@stat.osu.edu

› Principles of Neuroimaging

Martin A. Lindquist*, Johns Hopkins University

Understanding the brain is arguably among the most complex, important and challenging issues in science today. Neuroimaging is an umbrella term for an ever-increasing number of minimally invasive techniques designed to study the brain. These include a variety of rapidly evolving technologies for measuring brain properties, such as structure, function and disease pathophysiology. These technologies are currently being applied in a vast collection of medical and scientific areas of inquiry. This talk briefly discusses some of the critical issues involved in neuroimaging data analysis (NDA). This includes problems related to image reconstruction, registration, segmentation, and shape analysis. In addition, we will discuss various statistical models for conducting group analysis, connectivity analysis, brain decoding, and the analysis of multi-modal data. We conclude with discussing some open research problems in NDA.

EMAIL: mlindqui@jhsph.edu

› Treatment Selection Based on the Conditional Quantile Treatment Effect Curve

Xiao-Hua Andrew Zhou*, University of Washington

In this talk, we introduce a statistical framework for treatment assignment for a subgroup of patients, using their biomarker values based on casual inference. This new method is based on conditional quantile treatment effect (CQTE) curve, and CQTE curve's simultaneous confidence bands (SCBs), which can be used to represent the quantile treatment effect for a given value of the biomarker and select an optimal treatment for one particular patient. We then propose B-splines methods for estimating the CQTE curves and constructing simultaneous confidence bands for the CQTE curves. We derive the asymptotic properties of the proposed methods. We also conduct extensive simulation studies to evaluate finite-sample properties of the proposed simultaneous confidence bands. Finally, we illustrate the application of the CQTE curve and its simultaneous confidence bands in optimal treatment selection with a real-world data set. This is a joint work with Kaishan Han.

EMAIL: azhou@uw.edu

106. Genomics Integration

› Hierarchical Approaches for Integrating Various Types of Genomic Datasets

Marie Denis*, CIRAD, France
Mahlet G. Tadesse, Georgetown University

Advances in high-throughput technologies have led to the acquisition of various types of -omic data on the same biological samples. Each data type gives independent and complementary information that can explain the biological mechanisms of interest. While several studies performing independent analyses of each dataset have led to significant results, a better understanding of complex biological mechanisms requires an integrative analysis of different sources of –omic data. The proposed approach allows the integration of various genomic data types at the gene level by considering biological relationships between the different molecular features. Several scenarios and a flexible modeling, based on penalized likelihood approaches and EM algorithms, are studied and tested. The method is applied to genomic datasets from Glioblastoma Multiforme samples collected as part of the Cancer Genome Atlas project in order to elucidate biological mechanisms of the disease and identify markers associated with patients' survival.

EMAIL: marie.denis@cirad.fr

Inference Using Surrogate Outcome Data and a Validation Sample in EMR Study

Chuan Hong*, Harvard School of Public Health
Katherine Liao, Brigham and Women's Hospital
Tianxi Cai, Harvard School of Public Health

The link of genomic data to rich phenotypic data in Electronic Medical Records (EMR) systems has facilitated translational studies. Phenome-wide association studies (PheWAS) in EMR-linked cohort analyzes many phenotypes compared to a single genetic variant. One major challenge in PheWAS studies is the difficulty in efficiently characterizing accurate phenotypes with EMR. The ICD-9 code is often used as surrogates for the true outcome, which could result in misclassification and hampers the power of hypothesis tests. To deal with misclassification and to improve the efficiency, we propose a semi-supervised method using a small subset of data labeled from chart review. The proposed method allows multipe as well as various types of surrogates, and is robust to the model misspecification. Our simulation studies show that the proposed method provides better results than the existing methods under comparison. We conduct a data-analysis to investigate the association between the Low density lipoprotein genetic alleles with coronary artery disease in a rheumatoid arthritis cohort.

EMAIL: chong@hsph.harvard.edu

Fast and Accurate Batch Effect Adjustment for Non-Gaussian Genomic Data Using Generalized Linear Mixed Effects Model

Jun Chen*, Mayo Clinic
Xihong Lin, Harvard School of Public Health

Batch effects are pervasive in genomic data due to variability in sample handling. Without proper adjustment, batch effects can lead to reduced power to detect true associations and sometimes produce false associations due to confounding. Batch effects usually affect a large number of features simultaneously and therefore it is possible to borrow information across features. Previous methodology research on batch effect adjustment usually assumed the normality of the data, which may not always be met in practice. Here we propose a general framework based on generalized mixed effects model for batch effect adjustment for general outcomes such as DNA methylation beta-values. To overcome the computational hurdle, we developed an efficient algorithm based on penalized quasi-likelihood to estimate the model, which scales linearly with the sample size and the number of features. We conducted extensive simulations to assess the statistical properties of the proposed model. We found that our model usually resulted in higher power compared to existing methods while controlling for false positives. Finally, we illustrated our method with a real DNA methylation data set.

EMAIL: chen.jun2@mayo.edu

A Robust Approach in Differential Abundance Analysis Based on Longitudinal Metagenomic Sequencing Data

Lingling An*, University of Arizona
Dan Luo, University of Arizona

With the improvement of high-throughput sequencing technology, longitudinal metagenomic experiments have a great potential to reveal the impact of microbial features (e.g., species or genes) on human diseases or environmental factors. The counts in metagenomic data describe the relative abundance for a feature in the community. Such counts are usually with excessive zeros and highly skewed. For most cases the source of dispersion in the data is not clear, which means the (over)dispersion may not follow an assumed distribution. Thus, incorrect modeling the data could yield invalid inferences. Here we propose a distribution-free method that is implemented with functional response models for zero-inflated longitudinal metagenomic data by assessing the association between features and covariates. This model requires no distribution assumption of the data while also accounting for the correlation among repeated measurements by estimating a working correlation matrix. Through comprehensive simulation studies the performance of the proposed model is shown more robust compared with existing methods.

EMAIL: anling@email.arizona.edu

› iXplore: Software for Reproducible Interactive Data Visualization and Exploration

Zhicheng Ji*, Johns Hopkins University
Hongkai Ji, Johns Hopkins University

Data visualization and exploration are a key step towards insightful discoveries for most data-driven scientific research. Some existing software tools allow users to interactively visualize and explore data, such as enlarging a specific region of a large heatmap or scatter plot using the computer mouse. However, these software tools cannot record the details of data visualization and exploration procedures (e.g. the enlarged region of the heatmap) for others to reproduce the analysis. To address the problem, we develop iXplore, a software tool that performs interactive data visualization and exploration and can record and save the analysis procedures and results. The saved file can be loaded by other users on other computers to easily reproduce the whole data exploration procedures. iXplore provides convenient data visualization functions such as interactive and zoomable heatmaps and scatter plots. iXplore also provides some most commonly used statistical analysis functions such as clustering and differential analysis. iXplore is available online: https://zhiji.shinyapps.io/ixplore/.

EMAIL: zhichengji@gmail.com

107. Diagnostic and Screening Tests

› Novel Agreement Statistics Using Empirical Estimates of the Probability of Chance Agreement

Jarcy Zee*, Arbor Research Collaborative for Health
Laura H. Mariani, University of Michigan
Laura Barisoni, University of Michigan
Brenda W. Gillespie, University of Michigan

Many reproducibility measures correct for chance, typically by estimating the probability of chance agreement using marginal proportions of ratings observed by each rater. Some also make assumptions about the proportion of observations subject to random ratings. Estimates of agreement can therefore vary widely based on the statistic used. We propose novel agreement statistics for two raters and a binary rating that empirically estimate the probability of chance agreement by asking raters which observations or what proportion of observations were rated with uncertainty. We simulated data with various levels of prevalence, bias between raters, proportion of observations rated with uncertainty, number of observations rated, and amount of true agreement. We then tested the bias and efficiency of our proposed measures compared to existing measures, including Cohen's χ Scott's ϖ, Gwet's AC1, Bennett's S, and Krippendorff's χ Finally, we illustrate the use of these measures with data collected from pathologist ratings of structural features in renal biopsies in the Nephrotic Syndrome Study Network (NEPTUNE).

EMAIL: Jarcy.Zee@arborresearch.org

› Relationship Between Roe and Metz Simulation Model for Multireader Diagnostic Data and Obuchowski-Rockette Model Parameters

Stephen L. Hillis*, University of Iowa

For the typical diagnostic-radiology study design, each case undergoes several diagnostic tests and the resulting images are interpreted by several readers. The diagnostic tests are compared with respect to reader performance outcomes that are functions of the reader ROC curves. These reader-performance outcomes are frequently analyzed using the Obuchowski and Rockette method, which allows conclusions to generalize to both the reader and case populations. The Roe and Metz simulation model emulates confidence-of-disease data collected from such studies and has been used for evaluating reader-performance analysis methods. However, because the Roe and Metz model parameters are expressed in terms of a continuous decision variable rather than in terms of reader performance outcomes, it has not been possible to evaluate the realism of the RM model. I derive the relationships between the Roe-Metz and Obuchowski-Rockette model parameters for the empirical AUC reader-performance outcome. These relationships make it possible to evaluate the realism of the RM parameter models and to assess the performance of Obuchowski-Rockette parameter estimates.

EMAIL: steve-hillis@uiowa.edu

Back-End Disease Screening Using Information Theory

Xichen Mou*, University of South Carolina
Joshua M. Tebbs, University of South Carolina
Christopher S. McMahan, Clemson University
Christopher R. Bilder, University of Nebraska, Lincoln

Group testing is a strategy that pools individual specimens together and then tests the pools for the presence of a disease, such as HIV. When compared to testing individuals one by one, group testing can significantly reduce the number of tests needed to screen all of the individual specimens. However, one concern that arises in the implementation of group testing is diagnostic accuracy. Previous work that addressed improving accuracy involves performing additional confirmatory tests after initial diagnoses have been made. Borrowing complementary techniques from information theory, we investigate the potential use of coding designs to isolate specific individuals who should be targeted for retesting first. Coding theory problems have a similar structure to those problems which arise when investigating diagnostic accuracy in group testing. This talk will begin to merge these topics for possible use in large-scale disease testing applications.

EMAIL: xmou@email.sc.edu

A Bayesian Semiparametric Approach to Correlated ROC Surfaces with Constraints

Zhen Chen*, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health
Beomseuk Hwang, Chung-Ang University

In application of diagnostic accuracy, it is possible that a priori information may exist regarding the test score distributions, either between different disease populations for a single test or between multiple correlated tests. Motivated by a study on diagnosing endometriosis, we propose an approach to estimating diagnostic accuracy measures that can incorporate different stochastic order constraints on the test scores when an ordinal true disease status is in consideration. We show that the Dirichlet process mixture provides a convenient framework to both flexibly model the test score distributions and embed the a priori ordering constraints. We also utilize the Dirichlet process mixture to model the correlation between multiple tests. In taking a Bayesian perspective to inference, we develop an efficient Markov chain Monte Carlo algorithm to sample from the posterior distribution and provide posterior estimates of the receiver operating characteristic surfaces and the associated summary measures. The proposed approach is evaluated with extensive simulation studies, and is demonstrated with an application to the endometriosis study.

EMAIL: chenzhe@mail.nih.gov

A Hypothesis Testing Framework for Validating an Assay for Precision

Michael P. Fay*, National Institute of Allergy and Infectious Diseases, National Institutes of Health
Michael C. Sachs, National Cancer Institute, National Institutes of Health and Karolinska Institutet, Sweden
Kazutoyo P. Miura, National Institute of Allergy and Infectious Diseases, National Institutes of Health

A common way of validating an assay for precision is through the m:n:thetab procedure, where m levels of an analyte are measured with n replicates at each level, and if all m estimates of coefficient of variation (CV) are less than thetab, then the assay is declared validated for precision within the range of the m analyte levels. Two limitations of this procedure are: there is no clear statistical statement of precision upon passing, and it is unclear how to modify the procedure for assays with constant standard deviation. We reframe the m:n:thetab procedure as a set of m hypothesis tests. This reframing motivates the m:n:q procedure, which upon completion delivers a 100q% upper confidence limit on the CV. Additionally, for a post-validation assay output of y, the method gives an "effective standard deviation interval" of log(y) plus or minus r, which is a 68% confidence interval on log(mu), where mu is the expected value of the assay output for that sample. Further, the m:n:q procedure can be straightfowardly applied to constant standard deviation assays. We apply this new precision validating procedure to a growth inhibition assay.

EMAIL: mfay@niaid.nih.gov

› Notes on the Overlap Measure as an Alternative to the Youden Index: How are they Related?

Hani Samawi, Georgia Southern University
Jingjing Yin, Georgia Southern University
Haresh Rochani*, Georgia Southern University
Viral Panchal, Georgia Southern University

The ROC (Receiver Operating Characteristic) curve is frequently used for evaluating and comparing diagnostic tests. The Youden index, as one of the ROC summary index, measures the effectiveness of a diagnostic marker and enables the selection of an optimal threshold value (cut -off point) for the marker. Recently, the overlap coefficients (OVL), which capture the similarity between two distributions directly, are considered as alternative indices for determining the diagnostic performance of markers, that is, larger (smaller) overlap indicating poorer (better) diagnostic accuracy. This paper compare s the similarities and dissimilarities of the Youden index and OVL measures as well as the advantages of OVL measures over the Youden index in some situations. Numerical examples as well as a real data analysis with differentially expressed gene biomarkers are provided.

EMAIL: hrochani@georgiasouthern.edu

108. Methods for RNA-seq Data

› Biomarker Detection and Categorization in RNA-seq Meta-Analysis using Bayesian Hierarchical Model

Tianzhou Ma*, University of Pittsburgh
Faming Liang, University of Florida
George C. Tseng, University of Pittsburgh

Meta-analysis combining multiple transcriptomic studies increases statistical power and accuracy in detecting differentially expressed (DE) genes. As the NGS experiments become mature and affordable, more RNA-seq datasets are available in the public domain. A naive approach to combine multiple RNA-seq studies is to apply differential analysis to each study and then combine the summary statistics by conventional meta-analysis methods. Such a two-stage approach loses statistical power, especially for genes with short length or low expression abundance. In this paper, we propose a full Bayesian hierarchical model for RNA-seq meta-analysis by integrating information across studies and modeling potentially heterogeneous differential signals across studies via latent variables. A Dirichlet process mixture prior is further applied on the latent variables to provide categorization of detected biomarkers according to their DE patterns across studies, facilitating improved interpretation and biological hypothesis generation. Simulations and a real application on multi-brain-region HIV-1 transgenic rats demonstrate improved sensitivity, accuracy and biological findings of the method.

EMAIL: tim28@pitt.edu

› Unbiased Estimation of Parent-of-Origin Effects Using RNA-seq Data from Human

Vasyl Zhabotynsky*, University of North Carolina, Chapel Hill
Wei Sun, Fred Hutchinson Cancer Research Center
Kaoru Inoue, University of North Carolina, Chapel Hill
Terry Magnuson, University of North Carolina, Chapel Hill
Mauro Calabrese, University of North Carolina, Chapel Hill

RNA sequencing allows one to study allelic imbalance of gene expression, which may be due to genetic factors or genomic imprinting. It is desirable to model both genetic and parent-of-origin effects simultaneously to avoid confounding and to improve the power to detect either effect. In a study of experimental cross, separation of genetic and parent-of-origin effects can be achieved by studying reciprocal cross of two inbred strains. In contrast, this task is much more challenging for outbred population such as human population. To address this challenge, we propose a new framework to combine experimental strategies and novel statistical methods. Specifically, we propose to collect RNA-seq data from the children of family trios as well as phased genotype data from the trios, and we have developed a new statistical method to estimate both genetic and parent-of-origin effects from such data sets. We demonstrated this approach by studying 30 trios of HapMap samples. Our results support some of previous finding of imprinted genes and also recover some previously unknown imprinted genes.

EMAIL: vasyl@unc.edu

Power Calculation and Inference on RNA Sequencing Data

Xuechan Li*, Duke University
Jichun Xie, Duke University
Kouros Owzar, Duke University

We propose power and sample size calculation methodology for designing RNA high-throughput sequencing studies accounting for multiple testing within the false-discovery rate (FDR) framework. The underlying model assumes that conditional on the value of the trait, the numbers of reads mapped to each gene follows a negative-binomial distribution. The model allows for gene-specific mean and dispersion parameters. The inference is conducted on the basis of the asymptotic distribution of a Wald statistic. Our method can accommodate both binary and quantitative traits. We verify that the accuracy of the theoretical FDR and the false non-discovery rate (FNR) for our method through simulation studies. An R extension package is provided to implement the proposed power and sample-size calculation methodology to facilitate the design of RNA sequencing studies.

EMAIL: xuechan.li@duke.edu

An Optimal Test with Maximum Average Power for Time Course Data of Counts with Applications to RNA-seq Data

Meng Cao*, Colorado State University
Jay Breidt, Colorado State University
Wen Zhou, Colorado State University
Graham Peers, Colorado State University

Differential expression (DE) analysis for time course data from RNA-seq experiments has provided transformative insights in biology and biomedical studies. Existing methods focus on transformed data. By contrast, we propose a new test that directly accounts for both count data and temporal autocorrelation. Conditional on a latent linear Gaussian process, the data are modeled by negative binomial distributions. Motivated from Hwang and Liu (2003), a likelihood ratio statistic is further developed and employed for constructing the test while controlling the false discovery rate (FDR). The proposed test has the optimality property of maximum average power. Furthermore, while existing methods do not specifically address the testing of interactions between treatment and temporal dynamics, our test can detect DE genes either through a mean shift or through interactions. Simulation studies show that the proposed test has well-controlled FDR and good power. We applied our method to RNA-seq data from a study on temporal dynamics in photosynthesis, and discovered novel insights into the photosynthetic mechanism under different biological states.

EMAIL: meng.cao@colostate.edu

A Full Hierarchical Bayesian Model for Paired RNA-seq Data

Yuanyuan Bian*, University of Missouri, Columbia
Jing Qiu, University of Delaware
Zhuoqiong He, University of Missouri, Columbia

High-throughput RNA sequencing (RNA-seq) data measures the expression abundance through read counts of sequence and is able to detect differentially expressed genes (DEGs) under different conditions. Among those well-designed experiments for gene expression analysis, the paired design is an effective method to compare two conditions by controlling variability within the pair. Several methods have been proposed for paired RNA-seq analysis. However, most of them do not consider the heterogeneity in treatment effect among pairs that can naturally arise in real data, especially for human dataset. In addition, it has been reported in literature that the FDR control of many existing methods of identifying DE genes using RNA-seq data has been problematic. We present a full hierarchical Bayesian model for the paired RNA-seq count data that accounts for variation of treatment effect among pairs and controls the FDR through the posterior expected FDR. By simulation studies and real data analysis, we demonstrate that our approach is able to control FDR at the nominal level while all other competitive methods fail to and have slightly higher power under comparable actual FDR.

EMAIL: yb42c@mail.missouri.edu

Alternative Shrinkage Estimators for Effect Sizes in Analysis of RNA-seq Data

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

RNA-seq differential expression analysis often focuses on the generation of gene lists, while another aspect is the degree of accuracy of effect size estimates: the log fold changes associated with contrasts. To reduce highly variable effect sizes from low count genes, a common approach is to use gene filtering and pseudocounts. Alternatively, shrinkage estimators for effect size can be used, eliminating the need for filtering decisions early in the analysis pipeline. However, shrinkage estimation may result in undesirable bias of large effect sizes toward zero. Here, we use large RNA-seq datasets to examine the loss of sensitivity from standard filtering approaches and investigate the performance of novel shrinkage estimators for effect size in RNA-seq.

EMAIL: michaelisaiahlove@gmail.com

109. Variable Selection and Identification Methods

Detection of Differentially Methylated Regions with Mean and Variance Combined Signals

Ya Wang*, Columbia University
Andrew E. Teschendorff, University College London
Martin Widschwendter, University College London
Shuang Wang, Columbia University

DNA methylation plays an important role in cancer progression. Methylation levels between different groups were found to be different in both means and variances. Existing methods to detect differentially methylated regions (DMRs) consider correlated neighboring CpG sites but use only mean signals. Here we propose a new DMR detection algorithm that uses mean and variance combined signals. In simulation studies, we demonstrated the superior performance of the proposed algorithm than methods using one type of signals when true DMRs have both. Applications to DNA methylation data of breast cancer (BRCA) and kidney cancer (KIRC) from TCGA and BRCA from GEO suggest that the proposed algorithm identifies additional cancer-related DMRs that were missed by methods using only one type of signals. Replication analysis using the two BRCA datasets suggests that DMRs detected using variance signals are reproducible. Comparison between normal tissues adjacent to tumors and normal tissues from age-matched cancer-free women from the GEO BRCA data shows that the field defects (early epigenetic alteration) can only be detected using the proposed algorithm.

EMAIL: yw2453@columbia.edu

A General Parametric Regression Framework for Multiple-Infection Group Testing Data

Dewei Wang*, University of South Carolina
Juexin Lin, University of South Carolina

Group testing has been recognized as an efficient means of saving testing expenditures for large-scale screening practices. Recently, the use of multiplex assays has transformed its goal from detecting a single disease to multiple infections. Previous work on modeling multiple-infection group testing data focused on the estimation of disease prevalences where a multinomial model was assumed and individual level information was not considered. To incorporate such information, one can simply use a multinomial regression model, however, this model targets at joint inference among all the infections rather than marginal inference for each infection. To remedy this drawback, we propose a parametric regression framework which is able to jointly model multiple-infection group testing data and also to produce interpretable regression results including estimation, hypothesis testing, and variable selection for each infection separately. We illustrate the finite sample performance of our method through simulation and by applying it to chlamydia and gonorrhea screening data collected from the Infertility Prevention Project.

EMAIL: deweiwang@stat.sc.edu

Group Regularization for Zero-Inflated Regression Models with Application to HealthCare Demand in Germany

Shrabanti Chowdhury*, Wayne State University
Prithish Banerjee, Wayne State University
Saptarshi Chatterjee, Northern Illinois University
Broti Garai, West Virginia University
Himel Mallick, Harvard School of Public Health

In many practical applications, covariates have natural grouping structures, where variables in the same group are either mechanistically related or statistically correlated. Under such settings, variable selection should be conducted at both the group level and the individual variable level. Recently, various regularization methods have been extensively developed for variable selection in zero-inflated Poisson and zero-inflated negative binomial models. However, there is only limited methodological work on grouped variable selection for these zero-inflated count models. To bridge this gap, we specifically consider several popular group variable selection methods from the linear regression literature, such as the group lasso, group bridge, and sparse group lasso, and extend these techniques to zero-inflated count regression models. We investigate the finite sample performance of these methods through extensive simulation experiments and the analysis of health care demand in Germany. The open source software implementation of these methods is publicly available at: https://github.com/himelmallick/Gooogle.

EMAIL: gg0658@wayne.edu

Group SLOPE as Method for SNPs Selection with the False Discovery Rate Control

Damian Brzyski*#, Indiana University and Jagiellonian University, Poland
Alexej Gossmann, Tulane University
Weijie Su, University of Pennsylvania
Malgorzata Bogdan, University of Wroclaw, Poland

We extend the idea of Sorted L-One Penalized Estimation (SLOPE) to deal with the situation when one aims at selecting whole groups of explanatory variables instead of single regressors. We formulate the respective convex optimization problem, gSLOPE (group SLOPE), and propose an efficient algorithm for its solution. We also define a notion of the group false discovery rate (gFDR) and provide a choice of the sequence of tuning parameters for gSLOPE so that gFDR is provably controlled at a prespeciffied level if the groups of variables are orthogonal to each other. Moreover, we prove that our procedure adapts to unknown sparsity and is asymptotically mini- max. We also provide a method for the choice of the regularizing sequence when variables in different groups are not orthogonal but statistically independent and illustrate its good properties with computer simulations. Finally, we illustrate the advantages of gSLOPE in the context of Genome Wide Association Studies. R package grpSLOPE with implementation of our method is available on CRAN.

EMAIL: dbrzyski@iu.edu

Identifying Sparse Effects in Gene-Set Analysis

Yi Zhang*, Brown University
Yen-Tsung Huang, Institute of Statistical Science Academia Sinica
Zhijin Wu, Brown University

Gene-set analysis aims to identify the association between expression profiles of a group of genes with phenotypes of interest. Results on the gene set/pathway level are shown to have better interpretability and higher reproducibility across studies. Existing methods have mainly targeted gene sets with an enrichment of differential expression in the form of mean expression level between phenotypes, typically disease cases versus controls. However, in complex diseases, abnormal expression may not occur in the same gene in all patients, even if the same pathway is involved. We propose a new class of test statistics that allows the identification of gene sets whose expression profiles in the cases to deviate from the controls, even if the signal is sparse and discordant among the genes. In addition, a novel testing procedure is provided to accommodate joint testing on self-contained and competitive null hypotheses. We demonstrate that the new method has great power gain over existing methods when the signal is sparse and illustrate the application in real data using microarray data from a type II diabetes study.

EMAIL: yi_zhang@brown.edu

110. Causal Inference in Survival Analysis and Clinical Trials

Regression and Causal Models for Composite Endpoint

Lu Mao*, University of Wisconsin, Madison

Composite failure times are common endpoints in clinical studies, especially in cardiovascular trials. Instead of treating all component events indiscriminately regardless of severity, the win ratio statistic proposed by Pocock et al. (2012) gives due priority to fatal events, and has thus become a popular alternative to the traditional time-to-first-event analysis. However, the win ratio is restricted to two-sample comparison and does not adjust for covariates or potential confounding by pre-treatment variables. In this study, we consider natural extensions of the win ratio to the regression and causal settings. A common theme of our approaches, as is with the original win ratio, is pairwise comparion for adjudication of win/loss. For that purpose, we develop a general theory for estimating equations of a U-statistic structure and use it to derive asymptotic properties of the proposed estimators. Simulation studies show that the proposed methods perform well in finite samples. A dataset from a recent cardiovascular trial is analyzed to illustrate our methods.

EMAIL: lmao@unc.edu

Constructing a Confidence Interval for the Fraction Who Benefit

Emily J. Huang*, Johns Hopkins University
Ethan X. Fang, The Pennsylvania State University
Michael Rosenblum, Johns Hopkins University

In randomized trial analyses, the average treatment effect is commonly used to compare treatment with control. Another parameter of practical interest is the fraction who benefit from treatment. Formally, this is the proportion of patients whose potential outcome under treatment is better than that under control. Inference about this parameter is a challenging problem because the fraction who benefit is generally non-identifiable, though bounds are identifiable. Standard confidence interval methods, such as nonparametric bootstrap, can have poor performance for this problem. We propose a new method of constructing a confidence interval for the fraction who benefit, using randomized trial data. We consider the case of a binary or ordinal outcome. Our method does not require that any assumptions be made about the joint distribution of the potential outcomes. However, it can incorporate support restrictions on the joint distribution, such as the no harm assumption. Through simulation, we compare the proposed confidence interval procedure to existing methods with respect to coverage probability, average width, and computational efficiency.

EMAIL: ehuang19@jhu.edu

A Principal Stratification Approach to Evaluate the Causal Effect of Patient Activation Intervention on Bone Health Behaviors

Yiyue Lou*, University of Iowa
Michael P. Jones, University of Iowa
Stephanie W. Edmonds, University of Iowa
Fredric D. Wolinsky, University of Iowa

The Patient Activation after DXA Result Notification (PAADRN) study is a multi-center, pragmatic randomized clinical trial that was designed to improve bone health. 7749 participants were randomly assigned to either intervention group with usual care augmented by a tailored patient-activation DXA results letter accompanied by an educational brochure, or control group with usual care only. The primary analyses followed the standard intention-to-treat (ITT) principle, which might potentially underestimate the effect of intervention itself because PAADRN intervention might not have an effect if the patient did not read and remember the letter. In this paper, we apply principal stratification to evaluate the effectiveness of PADDRN intervention for subgroups, defined by the self-reported receipt of DXA result letter, while maintaining the constraints of randomization. We perform simulations to compare a new principal score weighting approach with the instrumental variable (IV) approach. We apply the methods to six outcome measures of bone health behaviors. Finally, we present sensitivity analyses to assess the effect of potential violations of relevant assumptions.

EMAIL: yiyue-lou@uiowa.edu

An Optimal Covariate-Adaptive Design to Balance Tiers of Covariates

Fan Wang*, The George Washington University
Feifang Hu, The George Washington University

Randomization is the 'gold standard' to estimate casual effects, but chance imbalance exists in covariate distribution among treatment groups. To address this issue, we propose a new covariate-adaptive design to improve the covariate balance. We specify an explicit definition of imbalance and control it by assigning units sequentially and adaptively. If covariates vary in importance, we partition them into tiers to ensure that important covariates have better balance. With a large number of covariates or a large sample size, our method has substantial advantages over traditional methods in terms of the covariate balance and computational time, and as such becomes an ideal technique in the era of big data. More crucially, our method attains the optimal covariate balance, in the sense that the estimated average treatment effect under our method attains its minimum variance asymptotically. All the above mentioned advantages of our method are further evidenced by extensive simulation studies.

EMAIL: wfanfp@gmail.com

Estimating Causal Effects from a Randomized Clinical Trial when Noncompliance is Measured with Error

Jeffrey A. Boatman*, University of Minnesota
David M. Vock, University of Minnesota
Joseph S. Koopmeiners, University of Minnesota
Eric C. Donny, University of Pittsburgh

Estimating the causal effect of a treatment in a randomized clinical trial is often complicated due to noncompliance. Existing methods to estimate the causal effect of a treatment assume that participants' compliance status is reported without error, but this is an untenable assumption when noncompliance is based on self-report. Biomarkers may provide more reliable indicators of compliance but cannot perfectly discriminate between compliers and noncompliers. However, by modeling the distribution of the biomarker as a mixture distribution and writing the probability of compliance as a function of the mixture components, we show how the probability of compliance can be directly estimated from the data even when compliance status is unknown. We develop a novel approach to estimate the causal effect that re-weights participants by the product of their probability of compliance given the biomarkers and the inverse probability of compliance given confounders. We show that our proposed estimator is consistent and asymptotically normal and demonstrate via simulation that the proposed estimator achieves smaller bias and greater efficiency than an ad hoc approach.

EMAIL: boat0036@umn.edu

Understanding Causal Effects of a Treatment on Survival in Observational Studies with Unmeasured Confounding

Fan Yang*, University of Chicago
Jing Cheng, University of California, San Francisco
Dezheng Huo, University of Chicago

Many clinical studies on survival outcomes based on observational data are challenged by unmeasured confounding. Instrumental variable (IV) methods are popular approaches to deal with unmeasured confounding and are increasingly being adopted in clinical studies. However, IV methods are not well developed for survival outcomes, especially for the Cox proportional hazards (PH) model which is the most popular regression model for censored survival data. Recently, there has been widespread use of the two stage residual inclusion (2SRI) method offered by Terza et al. (2008) for nonlinear models, and 2SRI has been the method of choice for analyzing PH model using IV in clinical studies. However, the causal parameter using 2SRI is only identified under a homogeneity assumption that goes beyond the assumptions of IV, and Wan et al. (2015) demonstrated that under standard IV assumptions, 2SRI could fail to consistently estimate the causal hazard ratio. In this paper, we develop an IV method to obtain a consistent estimate of the causal hazard ratio with a PH model specification under standard IV assumptions, and apply our method to an observational study of breast cancer.

EMAIL: fyang@health.bsd.uchicago.edu

111. Methods in Statistical Genetics

Rare Variant Higher Criticism for Sequencing Association Studies

Sixing Chen*, Harvard University
Xihong Lin, Harvard University

Whole genome sequencing analysis is challenged by rare variants. Traditional single SNP tests with rare variants are subject to poor power. Methods that test for association by aggregating the test statistics of multiple rare variants together in a genetic region are popular. These existing methods for rare variant analysis, such as SKAT, have good power when the signals are dense in the set of SNPs tested, but can have poor power when the signals are sparse. In contrast, thresholding methods for signal detection, such as Higher Criticism and Berk-Jones methods, have good power in the presence of sparse signals. However, they rely on the single SNP test statistics to be asymptotically normal. This normality assumption does not hold with rare variants for binary phentoypes and yield incorrect type I error. We propose a rare variant Higher Criticism approach to sparse signal detection that does not require normality of individual test statistics and accounts for correlation among the SNPs. The proposed test has higher power than aggregating methods, while allowing control of covariates and individual weighting of SNPs.

EMAIL: sixingchen@fas.harvard.edu

Sequential Test of Genetic Pleiotropy for Arbitrary Types of Traits

Daniel J. Schaid*, Mayo Clinic

Genetic pleiotropy -- the association of more than one trait with a genetic marker -- commonly occurs. Yet, most multivariate methods do not formally test pleiotropy. Rather, they test the null hypothesis that no traits are associated with a genetic marker; a statistically significant finding could result from only one trait driving the association. Furthermore, determining which traits are associated with a genetic marker, when traits are correlated, is often based on heuristic methods, To overcome these limitations, we derived a sequentially test of the null hypothesis that k+1 traits are associated, given that the null of k associated traits are was rejected. This provides a formal testing framework to determine the number of traits associated with a genetic variant, and which traits, while accounting for correlations among the traits. To allow for arbitrary types of traits, such as quantitative, binary, or ordinal, we used ideas from generalized linear models and generalized estimating equations to derive statistical test. Simulations show the properties of our methods, and application to an ongoing study of pleiotropy shows the utility of our approach.

EMAIL: schaid@mayo.edu

A Score Test for Genetic Class-Level Association with Non-Linear Biomarker Trajectories

Jing Qian*, University of Massachusetts, Amherst
Sara Nunez, Mount Holyoke College
Muredach P. Reilly, Columbia University
Andrea S. Foulkes, Mount Holyoke College

Emerging data suggest that the genetic regulation of the biological response to inflammatory stress may be fundamentally different to the genetic underpinning of the homeostatic control (resting state) of the same biological measures. We interrogate this hypothesis using a single-SNP score test and a novel class-level testing strategy to characterize protein coding gene and regulatory element-level associations with longitudinal biomarker trajectories in response to stimulus. Using the proposed Class-Level Association Score Statistic for Longitudinal Data (CLASS-LD), which accounts for correlations induced by linkage disequilibrium, the genetic underpinnings of evoked dynamic changes in repeatedly measured biomarkers are investigated. The proposed method is applied to data on two biomarkers arising from the Genetics of Evoked Responses to Niacin and Endotoxemia (GENE) Study. Our results suggest that the genetic basis of evoked inflammatory response is different than the genetic contributors to resting state, and several potentially novel loci are identified. A simulation study demonstrates appropriate control of type-1 error rates and relative computational efficiency.

EMAIL: qian@schoolph.umass.edu

Higher Criticism Test for Set-Based Mediation Effects of Multiple Genetic/Epigenetic Markers

Jincheng Shen*, Harvard School of Public Health
Xihong Lin, Harvard School of Public Health

It is of increasing interest to study the underlying mechanisms that whether the effect of environmental exposures on a disease outcome is mediated through genomic and epigenomic markers, e.g., DNA methylations. As multiple genomic markers in a genetic construct, e.g., CpG islands, genes, and pathways, are likely to function in concert, it is desired to test the effect of multiple mediators in a set. However, direct extensions of standard approaches of single mediator analysis could be conservative and thus severely underpowered due to possible sparse signals and the composite nature of the null hypothesis. We propose the Mediation Higher Criticism (MHC) tests for the mediation effects of multiple genomic/epigenetic markers in a genetic construct by allowing for sparse and weak signals and an arbitrary correlation among the markers. Simulation studies are performed to investigate the type I error rates and power of the proposed methods over a range of genetic regions with different correlation structures and signal sparsity. An application to the Normative Aging Study identifies DNA methylated regions that mediate the effects of smoking behavior on lung function.

EMAIL: jcshen@umich.edu

A Selective Region-Based Burden Test Identifies Most Susceptible Rare Variants with Improved Power and Interpretation in Sequencing Studies

Bin Zhu*, National Cancer Institute, National Institutes of Health
Nilanjan Chatterjee, Johns Hopkins University

In rare-variant association studies, aggregating low frequency variants within a region, such as a gene, increases statistical power. However, it's unclear which variants, or class of them, in a gene contribute most to the association. We proposed a region-based burden test (REBET) to simultaneously select and test significance of sub-regions demonstrating the strongest susceptibility. The sub-regions were predefined and shared common biologic characteristics, such as the protein domain or possible functional impact. Based on a subset-based approach considering local correlations between the combinations of test statistics of sub-regions, REBET was able to properly control the type I error rate, achieve higher power when rare variants clustered within sub-regions in the simulation studies and successfully identify associated protein domains in two case studies, while adjusting for multiple comparisons in a computational efficient fashion.

EMAIL: bin.zhu@nih.gov

Integrative Genomic Association Testing via Kernel Machine Mediation Analysis

Angela Hsiaohan Chen*, University of Illinois, Urbana-Champaign
Sihai Dave Zhao, University of Illinois, Urbana-Champaign

An integrative approach to association testing, which combines outcome and genotype data with other types of genomic information, has shown to be a more powerful approach to detect SNPs than the standard approach. Previously, Zhao et al. (2014) proposed a regression model for integrating genotype data, gene expression, and outcome, but their method required strong modeling assumptions on the relationship between expression and phenotype. We propose a method that can relax these assumptions by using a kernel machine (KM) regression framework that can allow for complex relationships, such as non-linear or interactive effects. Simulations and methodological comparisons demonstrate the benefits of our approach.

EMAIL: chen398@illinois.edu

Set-Based Tests Using the Generalized Berk-Jones Statistic in Genetic Association Studies

Ryan Sun*#, Harvard University
Peter Kraft, Harvard University
Xihong Lin, Harvard University

It has become increasingly popular to perform set-based inference on genetic markers instead of studying them individually. However, a significant challenge is that the associations between an outcome and markers in a set are expected to be sparse and weak, and existing methods may

not have the necessary power to detect such signals. Motivated by the Berk-Jones (BJ) statistic, which is notable for its strong asymptotic properties in detecting rare-weak signals, we propose a new test for association between a SNP-set and an outcome: the Generalized Berk-Jones (GBJ) statistic. Our GBJ statistic modifies the standard BJ to explicitly account for correlation between markers in a set, thus greatly increasing the power of the test when applied to correlated SNPs. We also provide a computationally efficient analytical p-value calculation for our method. The advantages of GBJ are demonstrated through rejection region analysis, simulation, and application to gene-level and pathway-level analyses of breast cancer data.

EMAIL: ryanrsun@gmail.com

112. Machine Learning And Massive Biological Data

› Replicates in High Dimensions, with Applications to Latent Variable Graphical Models

Kean Ming Tan*, Princeton University
Yang Ning, Cornell University
Daniela Witten, University of Washington
Han Liu, Princeton University

In classical statistics, much thought has been put into experimental design and data collection. In the high-dimensional setting, however, experimental design has been less of a focus. In this paper, we stress the importance of collecting multiple replicates for each subject in this setting. We consider learning the structure of a high-dimensional graphical model with latent variables, under the assumption that the latent variables take on a constant value across replicates within each subject. By collecting multiple replicates for each subject, we are able to estimate the conditional dependence relationships among the observed variables given the latent variables. To test the null hypothesis of conditional independence between two observed variables, we propose a pairwise decorrelated score test. We establish an inferential procedure to test whether there exists a hub node in the graph. Through numerical studies, we show that our proposal is able to estimate latent variable graphical models accurately compared to some existing proposals. We apply the proposed method to an fMRI data set obtained from multiple subjects while watching the television series Sherlock.

EMAIL: tan.keanming@gmail.com

› Higher-Order Supervised Dimension Reduction Methods: Applications to Electrocorticography

Genevera I. Allen*, Rice University
Frederick Campbell, Rice University
Michael Beauchamp, Baylor College of Medicine

Electrocorticography (ECoG) records electrical activity in the active human brain through hundreds of electrodes placed directly on the cortical surface of the brain. With its high temporal resolution, this technology is ideal for studying one of the most complex brain processes - human speech. Here, we seek to understand how the brain decodes multi-sensory stimuli to process speech. For speech decoding experiments, ECoG data can be organized as a massive four-dimensional tensor of speech trials by electrodes by time by frequency. We present novel higher-order supervised dimension reduction techniques that incorporate regularization to encourage sparsity in the electrodes and smoothness in the temporal and frequency domains. This approach yields interpretable patterns consisting of a sparse subset of electrodes and their smooth temporal and spectral firing patterns that best differentiate words from human speech. We demonstrate how our methods can be used for exploratory analysis, visualization, dimension reduction, and pattern recognition of massive tensor data produced by ECoG recordings studying speech decoding.

EMAIL: gallen@rice.edu

› Clustering with Hidden Markov Model on Variable Blocks for Single-Cell Data

Lynn Lin*, The Pennsylvania State University
Jia Li, The Pennsylvania State University

Advances in single-cell technologies have enabled high-dimensional measurement of individual cells in a high-throughput manner. A key first step to analyze this wealth of data is to identify distinct cell subsets from a mixed-population sample such as blood or tissue. In many

clinical applications, cell subsets of interest are often found in very low frequencies. This poses challenges for existing clustering methods. To address this issue, we propose a new mixture model, namely the Hidden Markov Model on Variable Blocks (HMM-VB). HMM-VB incorporates prior knowledge on the chain-like dependence structure among groups of variables, achieving the effect of dimension reduction as well as incisive modeling of the rare clusters. In a series of experiments on simulated and real data, HMM-VB outperforms other widely used methods.

EMAIL: llin@psu.edu

› An Integrative Statistical Framework for Multi-Modal Omics Data

George Michailidis*, University of Florida

It is becoming increasingly common for patients to be profiled across multiple molecular compartments -genomic, transcriptomic, proteomic, metabolomic, etc. We develop a framework that leverages recent developments in the estimation of high-dimensional multi-layered graphical models that provide insights on regulatory mechanisms across molecular compartments (layers), as well as on molecular interactions within each layer and are also capable of accommodating outcome variables such as disease risk, or patient survival times. We discuss algorithmic issues, establish theoretical properties of the estimates and apply them to real data from The Cancer Genome Atlas.

EMAIL: gmichail@ufl.edu

113. Extraordinary Possibilities for Mobile Health to Impact Precision Medicine

› Statistical and Dynamical Systems Modeling of Adaptive m-Intervention for Pain

Chaeryon Kang*, University of Pittsburgh
Daniel M. Abrams, Northwestern University
Jingyi Jessica Li, University of California, Los Angeles
Qi Long, Emory University
Nirmish R. Shah, Duke University

With the growing popularity of mobile phone technology, new opportunities have arisen for real-time adaptive medical intervention. The simultaneous growth of multiple "big data" sources allows for the development of personalized recommendations. In this project, we develop a new mathematical model for changes in subjective pain over time in patients with chronic conditions. The proposed model consists of a dynamical systems approach usingdifferential equations to forecast future pain levels, as well as a statistical approach tying system parameters to patient data (including reported pain levels, medication history, personal characteristics and other health records). The model is combined with statistical techniques to ultimately obtain optimized, continuously-updated treatment plans balancing competing demands of pain reduction and medication minimization. Application of the resulting personalized treatment plans to a currently active pilot study on mobile intervention in patients living with chronic pain due to sickle cell disease (SCD) will be presented.

EMAIL: crkang@pitt.edu

› Assessing Time-Varying Causal Effect Moderation in Mobile Health

Hyesun Yoo*, University of Michigan
Daniel Almirall, University of Michigan
Audrey Boruvka, University of Michigan
Katie Witkiewitz, University of New Mexico
Susan A. Murphy, University of Michigan

In mobile health (mHealth) for behavior change and maintenance, interventions are frequent and momentary. Typically, a great deal of information on patient states (e.g., stress), environmental factors, and behavioral responses is generated over time. Such intensive longitudinal intervention data is often collected by self-report or passively with the aid of sensors. One way in which such data may aid the design of a mobile intervention is the examination of effect moderation; i.e. inference about which factors strengthen or weaken the response to just-in-time interventions. Here, treatments, outcomes, and candidate moderators are time-varying. This paper (i) introduces a definition for moderated causal effects which is suitable for intensive longitudinal data, (ii) develops an easy-to-use weighted and centered regression approach for estimating these moderated effects (that can be used with

standard software), and (iii) compares the new method with standard longitudinal modeling. The method is illustrated using BASICS-Mobile, a smartphone-based intervention designed to curb heavy drinking and smoking among college students.

EMAIL: yoohs@umich.edu

Estimating Dynamic Treatment Regimes in Mobile Health Using V-Learning

Daniel J. Luckett*, University of North Carolina, Chapel Hill
Eric B. Laber, North Carolina State University
Anna R. Kahkoska, University of North Carolina, Chapel Hill
David M. Maahs, Stanford University
Elizabeth Mayer-Davis, University of North Carolina, Chapel Hill
Michael R. Kosorok, University of North Carolina, Chapel Hill

The vision for precision medicine is to use individual patient characteristics to inform a personalized treatment plan that leads to the best healthcare for each patient. Mobile technologies have a role to play as they offer a means to monitor a patient's health status in real-time and subsequently to deliver interventions if and when needed. Dynamic treatment regimes formalize individualized treatment plans as sequences of decision rules that map current patient information to recommended treatment. However, existing methods for estimating optimal treatment regimes are designed for a small number of fixed decision points on a coarse time-scale. We propose a new reinforcement learning method for estimating an optimal treatment regime that is applicable to data collected using mobile devices in an outpatient setting. The proposed method accommodates an indefinite time horizon and minute-by-minute decision making that is common in mobile health. We show the proposed estimators are consistent and asymptotically normal. The proposed methods are applied to estimate an optimal dynamic treatment regime for controlling blood glucose levels in patients with type 1 diabetes.

EMAIL: luckett@live.unc.edu

114. Integrative Analysis of Data From Different Sources

Subgroup Identification Based on Data from Multiple Sources

Menggang Yu*, University of Wisconsin, Madison

We consider a novel method for subgroup identification based on multiple individual patient level data sets. Our method is motivated by the situation when a single data set may lack power for robust subgroup identification. We assume that the multiple data sets share common influential features in subgroup identification and hence an integrated analysis of all the data sets would provide more power in variable selection and better performance in prediction, compared with separate analyses for each data set. The proposed method is based on a classification framework in subgroup identification. We incorporate weights and regularization to deal with possible heterogeneity among the data sets.

EMAIL: meyu@biostat.wisc.edu

Updating Established Prediction Models with New Biomarkers

Jeremy M.G. Taylor*, University of Michigan
Wenting Cheng, University of Michigan
Bhramar Mukherjee, University of Michigan

Models that give personalized predictions are abundant in the medical literature. There is a desire to improve these prediction models with new biomarkers. The information from an existing prediction model can be available in the form of coefficient estimates (with or without standard errors) or individual predicted probabilities. We investigate different approaches to incorporating such information while building a new prediction model that adds new biomarkers to the existing model. The situation is that these new biomarkers are measured on a small group of subjects while the existing model has been validated in large studies, but we do not have access to the data from the large studies. We formulate the problem in an inferential framework where the historical information is translated in terms of non-linear constraints on the

parameter space of the new model. We establish an approximate relationship between the regression coefficients in the two models. We develop both frequentists and Bayesian approaches. Simulation results suggest that the information from the established model can substantially improve the predictive power of the new model.

EMAIL: jmgt@umich.edu

› Integrated Analysis of Multidimensional Cancer Omics Data

Shuangge Ma*, Yale University

In recent cancer research, multidimensional studies have been extensively conducted, collecting multiple types of omics measurements on the same patients. The analysis of multidimensional data can lead to a deeper understanding of cancer biology and superior prediction models. However, such analysis is challenging with the high dimensionality, noisy nature of data, and, more importantly, the complex regulations among different types of omics changes. In our recent studies, we take advantage of developments in regularization techniques and network analysis, develop multiple novel analysis methods, and analyze data on the prognosis of melanoma and lung cancer. Such studies not only push forward system-based statistical analysis but also clinical utilization of omics signatures for multiple cancer types.

EMAIL: shuangge.ma@yale.edu

› Pathway and Gene Selection by the Integrative Analysis of Heterogeneous Omics Data

Quefeng Li*, University of North Carolina, Chapel Hill
Menggang Yu, University of Wisconsin, Madison
Sijian Wang, University of Wisconsin, Madison

In genetic studies, pooling information from multiplies studies is a common practice in order to identify biomarkers associated with diseases. The advance of genetic research discovers various sets of genes (referred as pathways) that function as a unit in the biological process. The existing integrative analysis methods do not build in such functional information, hence lack power in understanding the functions of genes. We propose a knowledge-based integrative analysis method for both pathway and gene identification. It combines the merits of both pathway analysis and integrative analysis and select both pathways and genes. The selected pathways are expected to provide more biological insights than individual genes. We demonstrate that the proposed method also brings statistical advantage. It only needs a fairly mild condition to asymptotically correctly identify the pathways. For completeness, the proposed method can also identify important genes under the selected pathways. Its finite-sample performance is shown to be superior than other methods in the simulation studies.

EMAIL: quefeng@email.unc.edu

115. Emerging Modeling Approaches in Large-Scale Data of Patients on Dialysis

› Issues in Profiling Dialysis Facilities

Danh V. Nguyen*, University of California, Irvine
Damla Senturk, University of California, Los Angeles
Yanjun Chen, University of California, Irvine
Luis Campos, Harvard University
Lishi Zhang, University of California, Irvine

Profiling or monitoring of health care providers, such as hospitals or dialysis facilities, with respect to a patient outcomes is an increasingly important activity at both state and federal levels. The typical patient outcomes assessed are overall or condition-specific mortality and 30-day (unplanned) hospital readmission. Profiling methods are based on random effects (REs) and fixed effects (FEs) hierarchical logistic regression models for a binary patient outcome. Limited profiling studies have not adequately elucidated the conditions under which REs and FEs are potentially effective. Towards this effort, we will report results on several comparative studies. Also, we explore novel follow-up analysis (post-profiling) to examine the extent to which facility-level factors (e.g., dialysis facility staffing level and composition) contribute to facility performance.

EMAIL: danhvn1@uci.edu

› Smoothing Spline Mixed-Effects Density Models for Clustered Data

Yuedong Wang*, University of California, Santa Barbara
Anna Liu, University of Massachusetts
Chi-Yang Chiu, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

We propose smoothing spline mixed-effects density models for the nonparametric estimation of density and conditional density functions with clustered data. The random effects in a density model introduce within-cluster correlation and allow us to borrow strength across clusters by shrinking cluster specific density function to the population average, where the amount of shrinkage is decided automatically by data. We present the construction of nonparametric mixed-effects density models for clustered data using smoothing spline ANOVA decompositions. Estimation is carried out using the penalized likelihood and computed using a Markov chain Monte Carlo stochastic approximation algorithm. We apply our methods to investigate evolution of hemoglobin density functions over time in response to guideline changes on anemia management for dialysis patients.

EMAIL: yuedong@pstat.ucsb.edu

Gateaux Differential-Based Boosting for Fitting Large-Scale Survival Data with Time-Varying Effects

Zhi He*, University of Michigan
Yanming Li, University of Michigan
Ji Zhu, University of Michigan
Yi Li, University of Michigan

Time-varying effects model is a flexible and powerful tool for modeling the dynamic changes of

covariate effects. In survival analysis, however, time-varying effects are often difficult to model. The computational burden increases quickly as the sample size or the number of predictors grows. which prohibits the application of existing statistical methods to large-scale data. In view of this gap, we propose a new Gateaux differential-based boosting procedure for simultaneously selecting and automatically determining the potential time-varying effects. Specifically, our procedure allows that in each boosting learning step only the best-fitting base-learner (and therefore the most informative covariate) is added to the predictor, and consequently encourages sparsity. In addition, our method controls smoothness, which is crucial for improving the predictive performance. The performance of the proposed method is examined by simulations and by applications to analyze national kidney dialysis data.

EMAIL: kevinhe@umich.edu

116. Current Trend in Topological Data Analysis

› Comparing Sparse Brain Activity Networks

Ben Cassidy*, Columbia University
DuBois Bowman, Columbia University

Functional magnetic resonance imaging of resting state brain activity has become a mainstay of modern neuroscience research. However, there are problems with existing methods for identifying, characterizing and comparing networks obtained from fMRI data, leading to many conflicting results in neuroimaging research. In this talk we introduce a new method for comparing networks, when the networks may be sparsely connected and of different sizes, by leveraging functional data analysis from statistics with persistent homology from mathematical topology.

EMAIL: b.cassidy@columbia.edu

› Topological Data Analysis on the Space of Brain Network Models

Moo K. Chung*, University of Wisconsin, Madison

Often used topological data analysis (TDA) deals with the Rips filtration on point cloud data, which builds a simplical complex. Subsequently, topological features such as barcodes and Betti-plots are extracted and used for statistical inference. However, the Rips filtration on point cloud data is very limiting as a statistical tool in various biomedical applications. In this talk, we present a new TDA framework, where a model with n parameters is treated as a point in an n-dimensional space endowed with a metric that measures the distance between two models with different parameters. In this abstract space, we build a filtration and construct topological features. The new framework avoids using a single optimal parameter that may

not be optimal in other studies and data sets. Instead of analyzing the model at the fixed parameter, we explore the dynamic topological changes of the model over different parameters. This new framework is applied in investigating the heritability of large-scale functional brain networks. This talk is based on preprint arXiv:1509.04771.

EMAIL: mkchung@wisc.edu

› OODA of Tree Structured Data Objects Using TDA

J.S. (Steve) Marron*, University of North Carolina, Chapel Hill

The field of Object Oriented Data Analysis has made a lot of progress on the statistical analysis of the variation in populations of complex objects. A particularly challenging example of this type is populations of tree-structured objects. Deep challenges arise, whose solutions involve a marriage of ideas from statistics, geometry, and numerical analysis, because the space of trees is strongly non-Euclidean in nature. Here these challenges are addressed using the approach of persistent homologies from topological data analysis. The benefits of this data object representation are illustrated using a real data set, where each data point is the tree of blood arteries in one person's brain. Persistent homologies gives much better results than those obtained in previous studies.

EMAIL: marron@unc.edu

117. Improving The Inferential Quality of Geospatial Data

› Confidence Intervals for Rate Ratios Between Geographic Units

Li Zhu*, National Cancer Institute, National Institutes of Health
Linda Pickle, StatNet Consulting
James Pearson, StatNet Consulting

Ratios of age-adjusted rates between a set of geographic units and the overall area are of interest to the general public and to policy stakeholders. The National Cancer Institute publishes cancer incidence and mortality rates and rate ratios on various geographic levels. These ratios are correlated due to two reasons – the first being that each region is a component of the overall area and hence there is an overlap between them; and the second in that there is spatial autocorrelation between the regions. Existing methods in calculating the confidence intervals of rate ratios take into account the first source of correlation. This talk addresses spatial autocorrelation in the rate ratios. The proposed method divides the rate ratio variances into three components, representing no correlation, overlap correlation, and spatial autocorrelation, representatively. Results with simulated and real cancer data show that with increasing strength and scales in spatial autocorrelation, the proposed method leads to substantial improvements over the existing method. If the data do not show spatial autocorrelation, the proposed method performs as well as the existing method.

EMAIL: li.zhu@nih.gov

› Physician Distribution Uncertainty in Three Common Workforce Data

Imam M. Xierali*, Association of American Medical Colleges (AAMC)
Marc A. Nivet, Association of American Medical Colleges (AAMC)

Accurate health workforce data are essential to effective workforce planning and policy. This study explored spatial uncertainty in physician practice location in three commonly used workforce data: AMA Masterfile (AMA), CMS NPI (NPI), and state licensure data (LIC). These data were triangulated to determine consistency in physician enumeration and distribution in Georgia. Physician addresses in two counties in Atlanta metro areas were linked to parcel land use classification. Residential address was assumed to be home address. Impact of home address on primary care spatial accessibility was examined. While more than 25% of Georgia's licensed physicians listed addresses outside the state, 92% of physicians in LIC could be matched to AMA and 90% to NPI. Location accuracy varied by physician gender, specialty, and underserved areas. In the two counties, 13.3% of NPI, 15% of AMA, and 24.9% of LIC addresses were found to be

residential. Impact of location uncertainty was stronger in underserved areas. These inconsistency and spatial uncertainty in physician datasets can limit projection accuracy of workforce supply and distribution.

EMAIL: ixierali@lsu.edu

› Optimal Regionalization for Multiscale Spatial Processes

Jonathan R. Bradley*, Florida State University
Christopher K. Wikle, University of Missouri
Scott H. Holan, University of Missouri

The presence of multiple spatial resolutions introduces challenging problems for uncertainty quantification in the spatial setting. In particular, the modifiable areal unit problem/ecological fallacy occur when modelling multiscale spatial processes. We investigate how these forms of spatial aggregation error can be quantified and used to develop a method for regionalization. Here, 'regionalization' refers to the problem of specifying areal units. We propose a criterion for spatial aggregation error, which is minimized to obtain an optimal regionalization. This criterion is developed theoretically through a new multiscale representation of the Karhunen–Loève expansion. Two examples are used to illustrate this method for regionalization: one using a federal dataset consisting of American Community Survey period estimates, and another analyzing environmental ocean winds.

EMAIL: bradley@stat.fsu.edu

› Mapping with(out) Confidence

David W. Wong*, George Mason University
Min Sun, George Mason University

Maps are often used to identify regions with extreme values (hot and cold spots). When values are assigned to different classes (colors) on a map, they are expected to be different. But if they are statistical estimates from surveys, they may be assigned to different classes even if they are statistically indifferent. When such maps are used in disease mapping, identified spatial patterns may not be real. Recently, the class separability concept was introduced as a map classification criterion to evaluate the likelihood that estimates in two classes are statistical different. Here, we demonstrate how the classification method can produce maps with reasonably separable classes. We also evaluated the performance of popular local cluster identification methods. Results show that existing methods do not produce consistent results and may leave out extreme values in delineating clusters. Identified clusters may include areas with values significantly different from values of other within-cluster units. Therefore, existing methods fail to define hot and cold spots comprehensively. We conclude that there is a lot of room to improve methods in disease cluster identification.

EMAIL: dwong2@gmu.edu

118. Meta-Analysis

› Gene- and Pathway-Based Association Tests for Multiple Traits with GWAS Summary Statistics

Il-Youp Kwak*, Lillehei Heart Institute, University of Minnesota
Wei Pan, Lillehei Heart Institute, University of Minnesota

To identify novel genetic variants associated with complex traits and to shed new insights on underlying biology, in addition to the most popular single SNP-single trait association analysis, it would be useful to explore multiple correlated (intermediate) traits at the gene- or pathway-level by mining existing single GWAS or meta-analyzed GWAS data. For this purpose, we present an adaptive gene-based test and a pathway-based test for association analysis of multiple traits with GWAS summary statistics. The proposed tests are adaptive at both the SNP- and trait-levels; that is, they account for possibly varying association patterns across SNPs and traits, thus maintaining high power across a wide range of situations. Furthermore, the proposed methods are general: they can be applied to mixed types of traits, and to Z-statistics or p-values as summary statistics obtained from either a single GWAS or a meta-analysis of multiple GWAS. Our numerical studies with simulated and real data demonstrated the promising performance of the proposed methods. The methods are implemented in R package aSPU available on CRAN.

EMAIL: ikwak@umn.edu

Bayesian Latent Hierarchical Model for Transcriptomic Meta-Analysis to Detect Biomarkers with Clustered Meta-Patterns of Differential Expression Signals

Zhiguang Huo*, University of Pittsburgh
Chi Song, The Ohio State University
George Tseng, University of Pittsburgh

Due to rapid development of high-throughput experimental techniques and fast dropping prices, many transcriptomic datasets have been generated and accumulated in the public domain. Meta-analysis combining multiple transcriptomic studies can increase statistical power to detect disease related biomarkers. In this paper, we introduce a Bayesian latent hierarchical model to perform transcriptomic meta-analysis. This method is capable of detecting genes that are differentially expressed (DE) in only a subset of the combined studies, and the latent variables help quantify homogeneous and heterogeneous differential expression signals across studies. A tight clustering algorithm is applied to detected biomarkers to capture differential meta-patterns that are informative to guide further biological investigation. Simulations and two examples including a microarray dataset from metabolism related knockout mice, and an RNA-seq dataset from HIV transgenic rats are used to demonstrate performance of the proposed method.

EMAIL: zhh18@pitt.edu

A Comparative Study for Rank Aggregation Methods with Focus on Genomic Applications

Xue Li*, Southern Methodist University
Xinlei Wang, Southern Methodist University

Rank aggregation, a meta-analysis method, combines different individual rank lists into one single rank list which is ideally more reliable. It has a rich history in the field of information retrieval, with applications to text mining, webpage ranking, meta-search engine building, etc. However, methods developed in such contexts are often ill suited for genomic applications, in which gene lists generated from individual studies are inherently noisy, due to various sources of heterogeneity. Further, because of missing or zero-count data, a portion of genes are not analyzed in all component studies, leading to partially ranked lists; and for some lists, only top-ranked genes are reported. In this study, we conduct a comprehensive comparison of existing methods for genomic applications with emphasis on partial and top-ranked lists. A systematic framework for classifying rank aggregation methods is proposed, performance of many popular methods is evaluated and practical guidelines for researchers who work in the field are provided.

EMAIL: xuel@smu.edu

Quantifying Publication Bias in Meta-Analysis

Lifeng Lin*, University of Minnesota
Haitao Chu, University of Minnesota

Publication bias (PB) is a serious problem in systematic reviews and meta-analyses that can affect the validity of conclusions. Currently, there are two classes of methods to handle PB: the selection models and the funnel-plot-based methods. The former class uses weight functions to adjust the overall effect size estimate, and they are usually employed as sensitivity analyses to assess PB. The latter class includes the regression and rank tests, and the trim and fill method. Although these methods have been widely used in applications, measures for quantifying PB are seldom studied in the literature. Such measures can serve as a characteristic of meta-analysis; also, they permit comparisons of PBs across different meta-analyses. Egger's regression intercept may be a candidate measure, but it lacks an intuitive interpretation. We introduce a new measure, the skewness of the standardized deviates, to quantify PB. It describes the asymmetry of the collected studies' distribution. Also, a new PB test is derived based on the skewness. Large sample properties of the new measure are studied, and its performance is illustrated using simulations and three case studies.

EMAIL: linl@umn.edu

Meta-Analysis of Rare Binary Events in Treatment Groups with Unequal Variability

Lie Li*, Southern Methodist University
Xinlei Wang, Southern Methodist University

In meta-analysis of rare binary events, zero counts arise frequently so that standard methods such as fixed-effect models

with continuity correction may cause substantial bias in estimation. Recently, Bhaumik et al. (2012) developed a simple average (SA) estimator based on a random-effects (RE) model. They proved that SA with the correction factor 0.5 (SA_0.5) is the least biased for large samples. However, RE models in previous work are restrictive because they assume that the variability in the treatment is equal to or always greater than that in the control. Under a general framework that allows groups with unequal variability but assumes no direction, we prove that SA_0.5 is still the least biased for large samples. Further, we consider the mean squared error (MSE) to assess estimation efficiency and show that SA_0.5 fails to minimize the MSE. Under a new RE model that accommodates groups with unequal variability, we compare the performance of various methods for both large and small samples via simulation, and draw conclusions about when to use which method in terms of bias and MSE. An example of rosiglitazone meta-analysis is used to provide further comparison.

EMAIL: liel@smu.edu

119. Statistical Genomics and Metabolomics

› Gene-Based Association Testing of Dichotomous Traits with Generalized Linear Mixed Models for Family Data

Chi-Yang Chiu*, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institute of Health
Ruzong Fan, Georgetown University
Yingda Jiang, University of Pittsburgh
Daniel Weeks, University of Pittsburgh
Momiao Xiong, University of Texas Health Science Center at Houston

We propose two approaches to test association between a dichotomous trait and genetic variants in a chromosome region for family-based data. The two approaches are based on additive generalized linear mixed models (GLMM) and generalized functional linear mixed models (GFLMM). The GLMM and GFLMM model the effect of a major gene as a fixed mean, the polygenic contributions as a random variation, and the correlation of pedigree members by kinship coefficients. The association between the dichotomous trait and the genetic variants is tested by testing the fixed mean to be 0 using likelihood ratio test (LRT) statistics. Simulation results indicate that the proposed LRT statistics control the type I error rates accurately and have higher power than the retrospective kernel and burden statistics developed by Schaid and his colleagues in most scenarios. The applications of our proposed methods to an age-related macular degeneration (AMD) family data set detect strong association between AMD and two known AMD susceptibility genes: CFH and ARMS2.

EMAIL: chiuchiyang@gmail.com

› A Bayesian Approach to Missing Data in Metabolomic Studies

Jasmit S. Shah*, University of Louisville
Shesh N. Rai, University of Louisville
Guy N. Brock, The Ohio State University
Aruni Bhatnagar, University of Louisville
Jeremy Gaskins, University of Louisville

Mass spectrometry in metabolomics makes it possible to measure the relative abundances of numerous metabolites in biological samples, which is useful to many areas of biomedical research. However, missing values (MVs) in metabolomics datasets are common and can arise due to both biological and technical reasons, such as an abundance below the limit of detection. Typically, such MVs are substituted by a minimum value, which may lead to inconsistent results in downstream analyses. In this study, we propose a Bayesian modeling approach to handling these MVs. We assume MVs are due to a mixture of two conditions: missingness due to truncation (missing not at random) and missingness for other reasons (assumed to be missing at random). The full complete data are assumed to be normally distributed with the observed data coming from the corresponding truncated normal. We propose an estimation scheme based on a data augmentation, Gibbs sampling algorithm. We conduct extensive simulations to investigate the inferential properties and the computational

capabilities of the proposed algorithm. We apply the method to impute MVs with applications to pre-clinical and clinical metabolomics studies.

EMAIL: jasmit.shah@louisville.edu

› Association Study of Children's Methylation and Growth Trajectory using Functional Mixed Models

Colleen McKendry, North Carolina State University
Arnab Maity*, North Carolina State University
Jung-Ying Tzeng, North Carolina State University
Cathrine Hoyo, North Carolina State University

Motivated from the Newborn Epigenetic STudy (NEST) data, we consider the problem of association study between children growth trajectory and children gene methylation profile while accounting for other confounders. We develop a functional semiparametric regression modeling framework where the response is functional variable (children growth trajectory measured over time), and scalar and vector valued covariates (gene methylation profile and other confounders). We model the joint effect of the gene methylation profile nonparametrically using the Gaussian process framework, and model the remaining confounders parametrically. We develop a hypothesis testing procedure for the effect of the gene methylation profiles using functional mixed effects models and variance components. We demonstrate our methodology via application to the NEST data.

EMAIL: amaity@ncsu.edu

› Make Big Data Alive: Interactive Data Visualization in Metabolomics Research

Jinxi Liu*, The George Washington University
Marinella Temprosa, The George Washington University

Metabolomics research has rapidly evolved. In this data-intensive field, effective data visualization tools empower researchers to present the big data in a meaningful way that people can quickly understand and use. Compared with traditional static graphics and tables, interactive visualization takes the concept further by allowing self-service faceting, probing and drill down. We developed several interactive data visualization applications for metabolomics research using Shiny by RStudio coupled with R packages ggvis and plotly. The applications present information including quality control and regression analysis of thousands of metabolites in different models. Results are conveyed both in data tables and statistical graphs. Users are allowed to view pointwise values using mouse-over controls, to drill down for detail through zooming, to compare and contrast the models and to display subsets or individuals by filtering and to download the data . The application can be published on websites to allow public or authenticated access and share with others. The above features enable a self-service, meaningful and flexible way to review and communicate data.

EMAIL: jeffljx12@gwu.edu

› Statistical Methods for Assessing Individual Oocyte Viability Through Gene Expression Profiles

Michael O. Bishop*, Utah State University
John R. Stevens, Utah State University
S. Clay Isom, Utah State University

In vivo derived oocytes are held as the gold standard for viability, other known origination methods are sub-par by comparison. Due to the low-viability of oocytes originating from these alternate methods, research was conducted to determine and quantify the validity of these alternate origination methods. However, the larger question of viability is on the individual oocyte level. We propose and compare methods of measurement based on gene expression profiles (GEPs) in order to assess oocyte viability, independent of oocyte origin. The first is based on a previously published wRMSD quantification of GEP differences. We also consider three novel methods: a distance comparison method, a tolerance interval method, and a classification-tree decision method; each utilizes a variable selection technique that focuses on the most differentially expressed genes. In our project, we obtain GEPs of individual swine oocytes and a general GEP distribution for in vivo oocytes. This distribution was the comparison standard for all oocytes, to gain a classi-

fication of viability. Each method is a valid method for driving viability decisions of the individual oocytes.

EMAIL: michael.o.bishop@aggiemail.usu.edu

› Predictive Metabolomic Profiling of Microbial Communities Using Shotgun Metagenomic Sequences

Himel Mallick*, Harvard School of Public Health, Broad Institute of MIT and Harvard University
Eric A. Franzosa, Harvard School of Public Health, Broad Institute of MIT and Harvard University
Lauren J. McIver, Harvard School of Public Health, Broad Institute of MIT and Harvard University
Soumya Banerjee, Harvard School of Public Health, Broad Institute of MIT and Harvard University
Alexandra Sirota-Madi, Harvard School of Public Health, Broad Institute of MIT and Harvard University
Aleksandar D. Kostic, Harvard School of Public Health, Broad Institute of MIT and Harvard University
Clary B. Clish, Broad Institute of MIT and Harvard University
Hera Vlamakis, Broad Institute of MIT and Harvard University
Ramnik Xavier, Massachusetts General Hospital, Harvard Medical School, Broad Institute of MIT and Harvard University
Curtis Huttenhower, Harvard School of Public Health, Broad Institute of MIT and Harvard University

Microbial community metabolomics, particularly in the human gut, are beginning to provide a new route to identify functions and ecology disrupted in disease. However, these data can be costly and difficult to obtain at scale, while shotgun metagenomic sequencing is readily available for populations of many thousands. Here we describe MelonnPan (Model-based Genomically Informed High-dimensional Predictor of Microbial Community Metabolic Profiles), a computational approach to predict metabolite pools given shotgun metagenomic information by incorporating biological prior knowledge in the form of microbial gene families or pathway abundance profiles. This allows us to accurately infer a community-wide metabolic network from a pool of up to one million microbial genes. Focusing on two independent gut microbiome datasets comprising over 150 patients with Crohn's disease, ulcerative colitis, and control participants, we demonstrate that our framework successfully recovers observed community metabolic trends in a large pool of microbial metabolites, including prediction of metabolic shifts associated with bile acid, fatty acids, steroids, prenol lipids, and sphingolipids.

EMAIL: hmallick@hsph.harvard.edu

120. Bayesian Methods For High-Dimensional and Clustered Data

› 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 are widely used and popular for analyzing time series datasets arising in economics, finance and genomics. A variety of Bayesian approaches for VAR models have recently been proposed in the literature. However, asymptotic properties of Bayesian VAR models in modern high-dimensional settings have not yet been investigated. In this paper, we examine the posterior convergence behavior for Bayesian vector autoregressive (VAR) models in a high-dimensional setting. In particular, 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 is smaller than n). In particular, this establishes posterior consistency under a variety of shrinkage priors, which introduce (group) sparsity in the columns of the coefficient matrix.

EMAIL: satyajitghosh90@ufl.edu

Bayesian Canonical Variate Regression with Incorporation of Pathway Information

Thierry Chekouo*, University of Minnesota, Duluth
Sandra E. Safo, Emory University

Recent advances in data collection and processing in biomedical research allow different data types to be measured on the same subjects, with each data type measuring different sets of characteristics, but collectively helping to explain underlying complex mechanisms. In some instances, phenotypic data are also available. The main goals of these problems are to study the overall dependency structure among the data types, and to develop a model for predicting future phenotypes. Canonical correlation analysis is oftentimes used for such problems. We present a Bayesian canonical correlation framework that simultaneously models the overall association between data types using only relevant variables, while also predicting future outcomes using the canonical correlation variates. In addition, through prior distributions, we incorporate in our model prior structural information (such as biological networks) within each data type that allows us to select functionally meaningful networks involved in the determination of canonical correlation variates. We demonstrate the effectiveness of the proposed approach using simulations and observed data.

EMAIL: ssafo@emory.edu

A Mixture Approach to Estimating a Population Value

Haoyu Zhang*, Johns Hopkins Bloomberg School of Public Health
Tom Louis, Johns Hopkins Bloomberg School of Public Health

When analyzing hierarchical (clustered) data, the analyst needs to control the weights used in combining. The target of inference is a population level feature computed from the cluster-specific features. Specifically, variance minimizing weights or more general use of a population-level likelihood can produce an off-target (large sample biased) estimate. We hereby identify situations where a two-level, hierarchical modeling approach is appropriate, while taking into account the relationship between underlying parameters and sample size. The hierarchical modeling approach is based on either a parametric or a non-parametric mixing distribution. When the true distribution is a subset of the assumed form, the hierarchical modeling approach will produce properly targeted estimate of population-level parameters using all stratum-specific data. More generally, the approach can be used to make inference for a permutation invariant functional of unit specific parameters. Our proposed method will be demonstrated using a real-life dataset.

EMAIL: andrew.haoyu@gmail.com

High Dimensional Confounding Adjustment Using Continuous Spike and Slab Priors

Joseph Antonelli*, Harvard School of Public Health
Giovanni Parmigiani, Harvard School of Public Health
Francesca Dominici, Harvard School of Public Health

In many studies, interest lies in the effect of an exposure on an outcome. Valid estimation of an exposure effect requires proper controlling of confounding. If the number of covariates is large relative to the number of observations in the study, then direct control of all covariates is infeasible. In such cases, variable selection or penalization can reduce the dimension of the covariate space in a manner that allows us to control for confounding. In this article, we propose continuous spike and slab priors when the number of covariates P exceeds the number of observations N. We construct the probability of a parameter being included in the 'slab' component of the prior to increase the probability of including confounders into the model. We illustrate our approach using both full posterior exploration via MCMC as well as posterior mode estimation via coordinate descent. We use theoretical arguments and empirical studies to illustrate how our prior reduces shrinkage for important confounders and leads to improved estimates of treatment effects.

EMAIL: jantonelli111@gmail.com

Scalable Bayesian Nonparametric Learning for High-Dimensional Cancer Genomics Data

Chiyu Gu*, University of Missouri
Subha Guha, University of Missouri
Veera Baladandayuthapani, University of Texas MD Anderson Cancer Center

'Omics datasets, which involve intrinsically different sizes and scales of high-throughput data, offer genome-wide, high-resolution information about the biology of lung cancer. One of the main goals of analyzing these data is to identify differential genomic signatures among samples under different treatments or biological conditions. e.g., treatment arms, tumor (sub)types, or cancer stages. We construct an encompassing class of nonparametric models called PDP-Seq that are applicable to mixed, heterogeneously scaled datasets. Each platform can choose from diverse parametric and nonparametric models including finite mixture models, finite and infinite hidden Markov models, Dirichlet processes, and zero and first order PDPs, which incorporate a wide range of data correlation structures. Simulation studies demonstrate that PDP-Seq outperforms many existing techniques in terms of accuracy of genomic signature identification. The pathway analysis identifies upstream regulators of many genes that are common genetic markers in multiple tumor cells.

EMAIL: cgz59@mail.missouri.edu

High-Dimensional Graph Selection and Estimation Consistency for Bayesian DAG Models

Xuan Cao*, University of Florida
Kshitij Khare, University of Florida
Malay Ghosh, University of Florida

Covariance estimation and selection for high-dimensional multivariate datasets is a fundamental problem in modern statistics. Gaussian directed acyclic graph (DAG) models is a popular class of models used for this purpose. Gaussian DAG models introduce sparsity in the Cholesky factor of the inverse covariance matrix, and the sparsity pattern in turn corresponds to specific conditional independence assumptions on the underlying variables. A variety of priors have been developed in recent years for Bayesian inference in DAG models, yet crucial convergence and sparsity selection properties for these models have not been thoroughly investigated. In this paper, we consider DAG-Wishart priors, which is a flexible and general class of priors for Gaussian DAG models with multiple shape parameters. Under mild regularity assumptions, we establish strong graph selection consistency and establish posterior convergence rates for estimation when the number of variables p is allowed to grow at an appropriate sub-exponential rate with the sample size n.

EMAIL: caoxuan@ufl.edu

Adaptive Bayesian Spectral Analysis of Nonstationary Biomedical Time Series

Scott A. Bruce*#, Temple University
Martica H. Hall, University of Pittsburgh
Daniel J. Buysse, University of Pittsburgh
Robert T. Krafty, University of Pittsburgh

Many studies of biomedical time series aim to measure the association between frequency-domain properties of time series and clinical covariates. However, the time-varying dynamics of these associations are largely ignored due to a lack of methods that can assess the changing nature of the relationship in time. This article introduces a method for the automatic analysis of the association between the time-varying power spectrum and covariates. The procedure adaptively partitions the grid of time and covariate values into an unknown number of approximately stationary blocks and estimates local spectra within blocks through penalized splines. The approach is formulated in a Bayesian framework, where the number and locations of partition points are random, and fit using reversible jump Markov chain Monte Carlo techniques. Inference averaged over the distribution of partitions allows for analysis of spectra with both smooth and abrupt changes. The proposed methodology is used to analyze the association between the time-varying spectrum of heart rate variability and self-reported sleep quality in a study of older adults serving as the primary caregiver for their ill spouse.

EMAIL: scott.bruce@temple.edu

121. Bayesian Methods and Genetics/Genomics

Cross-Species Gene Expression Analysis: Resemblance between Human and Mice Models

Md Tanbin Rahman*, University of Pittsburgh
Tianzhou Ma, University of Pittsburgh
George Chien-Cheng Tseng, University of Pittsburgh

Mice models have been widely used in medical research to delve into the mechanism in which the human body works. However, the use of mice model in transcriptomic studies remains controversial. Two recent papers on PNAS discussed the effectiveness of mice models and reached two contradictory conclusions on whether genomic responses in mice models mimic human inflammatory diseases, using the same datasets. Motivated by that, we here propose a robust statistical methodology to quantify how much the genomic response of a mouse model may resemble human disease. We define both a global resemblance score and pathway specific resemblance scores to assess the overall resemblance of mice model as well as the resemblance of the mice model in a certain pathway. The scores are estimated using a Bayesian hierarchical model and adjusted by the background level of resemblance under null. We have applied our methods to the inflammatory disease datasets with sepsis, trauma mice models and sepsis,trauma human models and found a high resemblance score between human and mice models for immune-related pathways which is reasonable.

EMAIL: MDR56@pitt.edu

A MAD-Bayes Algorithm for State-Space Inference and Clustering with Application to Different ChIP-seq Problems

Kailei Chen*#, University of Wisconsin, Madison
Chandler Zuo, University of Wisconsin, Madison
Sunduz Keles, University of Wisconsin, Madison

Analyses of Chromatin immunoprecipitation followed by sequencing (ChIP-seq) data typically concern two problems: i) binding state inference, which aims at detecting the DNA loci occupied by the transcription factor of interest; and ii) allele-specific binding analysis which incorporates the heterozygous Single Nucleotide Polymorphisms to study how single base pair variations between the two strands impact binding events. Matrix Based Analysis for State-space Inference and Clustering (MBASIC) framework is the first method to encompass joint analysis of multiple ChIP-seq datasets for different models in a unified framework. The EM based estimation structure of this framework hinders its applicability with large numbers of loci and samples. We address this limitation by developing a MAP-based Asymptotic Derivations from Bayes (MAD-Bayes) method under strong assumptions in MBASIC. This results in a K-means-like optimization algorithm which converges rapidly.

EMAIL: kchen@stat.wisc.edu

A Novel Region-Based Bayesian Approach for Genetic Association with Next Generation Sequencing (NGS) Data

Laurent Briollais*, Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital
Jingxiong Xu, University of Toronto

The discovery of rare variants through NGS is becoming a very challenging issue in human genetics. Because rare variants occur too infrequently in the general population, single-variant association tests lack power. We propose here a novel region-based statistic based on a Bayes Factor (BF) to assess evidence of association between a set of rare variants located on same chromosomal region and a disease outcome. Marginal likelihoods are computed under the null and alternative hypotheses assuming a binomial distribution for the rare variants count in the region and either a beta prior distribution or a mixture of Dirac and beta distribution for the probability of observing a rare variant at a specific locus. The hyper parameters are determined empirically from the data. A permutation test is used to assess the distribution of the BF under the null hypothesis. Our simulations studies showed that the new BF statistic outperforms popular methods such as the Burden test and SKAT under most situations. Our real application to a study of lung cancer including 262 cases and 260 matched controls suggests interesting genes such as TERT associated with the cancer outcome.

EMAIL: laurent@lunenfeld.ca

A Bayesian Hierarchical Model for Pathway Analysis with Simultaneous Inference on Pathway-Gene-SNP Structure

Lei Zhang*, University of Texas, Dallas
Swati Biswas, University of Texas, Dallas
Pankaj Choudhary, University of Texas, Dallas

Pathway analysis jointly tests the combined effects of all single nucleotide polymorphisms(SNPs) in all genes belonging to a molecular pathway. It is usually more powerful than single-SNP analyses if multiple associated variants of modest effects exist in a pathway. We develop a Bayesian hierarchical model that fully models the natural three level hierarchy, namely SNP-gene-pathway, unlike many other methods using ad hoc ways of combining such information. The joint modeling allows detecting not only the associated pathways but also testing for association with genes and SNPs within significant pathways and genes in a hierarchical manner. Appropriate priors are used to regularize the effects and hierarchical FDR is used for multiplicity adjustment of the entire inference procedure. To study the proposed approach, we conducted simulations with samples generated under realistic linkage disequilibrium patterns obtained from the HapMap project. Our method has higher power than some standard approaches for identifying pathways with multiple modest-sized variants. Moreover, in some settings, it can detect associated genes and SNPs, a feature unavailable in other methods.

EMAIL: lxz096120@utdallas.edu

PairClone: A Bayesian Subclone Caller Based on Mutation Pairs

Tianjian Zhou*, University of Texas, Austin
Peter Mueller, University of Texas, Austin
Subhajit Sengupta, NorthShore University HealthSystem
Yuan Ji, NorthShore University HealthSystem and University of Chicago

Tumor cell populations can be thought of as being composed of homogeneous cell subpopulations, with each subpopulation being characterized by overlapping sets of single nucleotide variants (SNVs). Such subpopulations are known as subclones and are an important target for precision medicine. Reconstructing such subclones from next-generation sequencing (NGS) data is one of the major challenges in precision medicine. We present PairClone as a new tool to implement this reconstruction. The main idea of PairClone is to model short reads mapped to pairs of proximal SNVs. In contrast, most existing methods use marginal reads for unpaired SNVs. Using Bayesian nonparametric models, we estimate posterior probabilities of the number, genotypes and population frequencies of subclones in one or more tumor sample. We use the categorical Indian buffet process (cIBP) as a prior probability model for subclones that are represented as vectors of categorical matrices that record the corresponding sets of mutation pairs. Performance of PairClone is assessed using simulated and real datasets. An open source software package can be obtained at http://www.compgenome.org/pairclone.

EMAIL: tjzhou95@gmail.com

Pathway-Based Integrative Bayesian Modeling of Multiplatform Genomics Data

Elizabeth J. McGuffey*, United States Naval Academy
Jeffrey S. Morris, University of Texas MD Anderson Cancer Center
Raymond J. Carroll, Texas A&M University
Ganiraju C. Manyam, University of Texas MD Anderson Cancer Center
Veerabhadran Baladandayuthapani, University of Texas MD Anderson Cancer Center

The identification of gene pathways involved in cancer development and progression and characterization of their activity in terms of multiplatform genomics can provide information leading to discovery of new targeted medications. Such drugs have the potential to be used for precision therapy strategies that personalize treatment based on the biology of an individual patient's cancer. We propose a two-step model that integrates multiple genomic platforms, and gene pathway membership information, to simultaneously identify genes significantly related to a clinical outcome, identify genomic platform(s) regulating each important gene, and rank pathways by importance to clinical outcome. We utilize a hierarchical Bayesian model with multiple levels of shrinkage priors to achieve efficient estimation, and our integrative framework allows us not only to identify important pathways and important genes within pathways, but also to gain insight as to the platform(s) driving the effects mechanistically. We apply our method to a subset of The Cancer Genome Atlas' publicly

available glioblastoma multiforme data set and identify potential targets for future cancer therapies.

EMAIL: elizabeth.mcguffey@gmail.com

122. Health Services Research

› Statistical Modeling for Heterogeneous Populations with Application to Hospital Admission Prediction

Jared D. Huling*, University of Wisconsin, Madison
Menggang Yu, University of Wisconsin, Madison

This work is motivated by risk modeling for large hospital and health care systems that provide services to diverse and complex patients. Modeling such heterogeneous populations poses many challenges. Often, heterogeneity across a population is determined by a set of factors such as chronic conditions. When these stratifying factors result in overlapping subpopulations, it is likely that the covariate effects for the overlapping groups have some similarity. We exploit this similarity by imposing constraints on the importance of variables. Our assumption is that if a variable is important for a subpopulation with one of the chronic conditions, then it should be important for the subpopulation with both conditions. However a variable can be important for the subpopulation with two particular chronic conditions but not for the subpopulations with just one. This type of assumption is reasonable and aids in borrowing strength across subpopulations. We prove an oracle property for our estimator and demonstrate its impressive performance in numerical studies and on an application in hospital admission prediction for the Medicare population of a large health care provider.

EMAIL: huling@wisc.edu

› Latent Class Dynamic Mediation Model with Application to Smoking Cessation Data

Jing Huang*, University of Pennsylvania
Ying Yuan, University of Texas MD Anderson Cancer Center
David Wetter, University of Utah

Traditional mediation analysis assumes that a study population is homogeneous and the mediation effect is constant over time, which may not hold in some applications. Motivated by smoking cessation data, we propose a latent class dynamic mediation model that explicitly accounts for the fact that the study population may consist of different subgroups and the mediation effect may vary over time. We use a proportional odds model to accommodate the subject heterogeneities and identify latent subgroups. Conditional on the subgroups, we employ a Bayesian hierarchical nonparametric time-varying coefficient model to capture the time-varying mediation process, while allowing each subgroup to have its individual dynamic mediation process. A simulation study shows that the proposed method has good performance in estimating the mediation effect. We illustrate the proposed methodology by applying it to analyze smoking cession data.

EMAIL: jing14@mail.med.upenn.edu

› Analysis of Meals on Wheels Recipients through Linkage to Electronic Health Records

Mingyang Shan*, Brown University
Roee Gutman, Brown University

Record linkage is a powerful tool that allows us to merge records from the same individual from multiple databases to use in a greater overall analysis. Due to data confidentiality issues, it is often the case that a unique identifying variable is not available to link the individuals, and the use of quasi-identifying information is necessary instead. The task becomes increasingly more complicated as databases become larger and the number of quasi-identifying variables is few. Motivated by the linkage of elderly Meals on Wheels recipients to their Medicare claims records, we examine a two stage method which identifies the overlapping individuals among two sources of data through record linkage, and subsequently selects control individuals based on electronic health records with similar medical history to use in comparative causal analyses.

EMAIL: mingyang_shan@brown.edu

Assessing the Benefits of Multiple Incompatible Donors for Transplant Candidates in Kidney Paired Donation

Mathieu Bray*, University of Michigan
Wen Wang, University of Michigan
Peter X-K. Song, University of Michigan
John D. Kalbfleisch, University of Michigan

In kidney paired donation (KPD), a transplant candidate with an incompatible donor joins a network of such pairs in an effort to reveal new transplant opportunities. The aim in KPD is to maximize the number of transplants achieved via donor exchange cycles among the pairs, as well as chains originating from altruistic donors. Previous literature has examined the benefits of accounting for probabilities of exchange failure (e.g. due to withdrawal, reversal of presumed compatibility, etc.), and fallback options in the case of failure (Li 2014, Bray 2015). In reality, a candidate may have several donors willing to join KPD, introducing new potential transplant opportunities, and possible immediate alternatives in case of failure. Our aim here is threefold: (i) we extend previous KPD optimization algorithms to allow candidates with multiple incompatible donors; (ii) through simulation, we evaluate the benefits, in terms of waiting time and graft survival for candidates, and total realized transplants within the network; (iii) we develop software that provides an interactive display of the KPD network while accounting for the various clinical features outlined above.

EMAIL: braymath@umich.edu

Measuring the Effects of Time-Varying Medication Adherence on Longitudinally Measured Health Outcomes

Mark Glickman*, Harvard University
Luis Campos, Harvard University

One of the most significant barriers to disease management is patients' non-adherence to their prescription medication. Quantifying the impact of medication non-adherence can be difficult because a patient's adherence may be changing over time. With the availability of detailed adherence data derived from electronic pill-top monitors, it is now possible to measure the effects of time-varying adherence on health outcomes. We present a modeling framework for patient outcomes from electronic monitored medication adherence data. The model assumes two ideal states for each patient; one in which a patient is perfectly adherent to a medication, and the other in which a patient in perfectly non-adherent. The mean outcome process varies dynamically between these two extremes as a function of the time-varying medication process. The framework permits the inclusion of baseline health characteristics, allows for missing adherence data, and can account for different medications, dosages and regimens. We demonstrate the modeling approach to a cohort of patients diagnosed with hypertension who were prescribed anti-hypertensive medication placed in electronic monitoring devices.

EMAIL: glickman@fas.harvard.edu

Development of a Common Patient Assessment Scale Across the Continuum of Care: A Nested Multiple Imputation Approach

Chenyang Gu*, Harvard University
Roee Gutman, Brown University

The assessment of patients' functional status across the continuum of post-acute care requires a common assessment tool. Different assessment tools are implemented in different health care settings and they cannot be easily contrasted. For patient leaving rehabilitation facilities, we equate the Minimum Data Set that is collected for all patients who stay in skilled nursing facilities and the Outcome and Assessment Information Set that is collected if they choose home health care provided by home health agencies. We consider equating as a missing data problem, and propose a two-stage procedure that combines nested multiple imputation and the multivariate ordinal probit model to obtain a common patient assessment scale across the continuum of care by imputing unmeasured assessments at multiple assessment dates. This procedure enables comparison of the rates of functional improvement experienced by patients treated in different health care settings using a common measure. The proposed procedure is compared to existing methods in a simulation study, and is illustrated using a real data set. Our procedures are general and can be used in other settings as well.

EMAIL: gu@hcp.med.harvard.edu

123. Neuroimaging

Discriminating Sample Groups with Multi-Way Biochemical Neuroimaging Data

Tianmeng Lyu, University of Minnesota
Eric Lock, University of Minnesota
Lynn E. Eberly*, University of Minnesota

High-dimensional linear classifiers, such as support vector machines (SVM) and distance weighted discrimination (DWD), are used to distinguish groups based on a large number of features. However, their use is limited to applications where a vector of features is measured per subject. In practice, data may be multi-way: measured over multiple dimensions, for example, metabolite abundance over multiple tissues, or gene expression over multiple time points. We propose a framework for linear classification of high-dimensional multi-way data, in which coefficients can be factorized into weights for each dimension. More generally, the coefficients for each measurement in a multi-way dataset have low-rank structure. This work extends existing classification techniques, and we have implemented and compared them to competing classifiers. We describe simulation results, and apply multi-way DWD to a neuroimaging study using magnetic resonance spectroscopy data over multiple brain regions to compare patients with and without spinocerebellar ataxia. Our method improves performance and simplifies interpretation over naive applications of full rank linear classification to multi-way data.

EMAIL: leberly@umn.edu

A Statistical Method for Advancing Neuroimaging Genetics

Xuan Bi*, Yale University
Heping Zhang, Yale University

Analyzing neuroimaging and genomic data is critical to the understanding of brain function. With the advent of modern technology, it has become feasible to collect both neuroimaging and genomic data from large-scale studies, such as the Philadelphia Neurodevelopmental Cohort and the Pediatric Imaging, Neurocognition, and Genomics Study. Utilization of such big data becomes a bottleneck that desperately needs to be resolved. In particular, insufficient work has been done through incorporating neuroimaging and genomic data simultaneously, which is known to be challenging. Such analysis includes large volumes of complex information, which is an emerging need and more daunting than analyzing typical neuroimaging data or genomic data. We develop an innovative statistical method that better defines the phenotypes by considering both neuroimaging and clinical data and can better combine generic variants. This strategy maximizes the power to identify genetic variants that have the biological implications through brain function and improve the quality of genetic studies over those based on the use of clinical diagnosis alone as the phenotypes.

EMAIL: xuan.bi@yale.edu

Generalized Mahalanobis Depth in Point Process and its Application in Neural Coding

Shuyi Liu*, Florida State University
Wei Wu, Florida State University

In this paper, we propose to generalize the notion of depth in temporal point process observations. The new depth is defined as a weighted product of two probability terms: 1) the number of event in each process, and 2) the center-outward ranking on the event times conditioned on the number of events. In this study, we adopt the Poisson distribution in the first term, and the Mahalanobis depth in the second term. We propose an efficient bootstrapping approach to estimate parameters in the defined depth. In the case of Poisson process, the observed events are order statistics, and the parameters can be estimated robustly with respect to sample size. We demonstrate the use of the new depth by ranking realizations from a Poisson process. We also test the new method in classification problems in simulations as well as real experimental data. It is found that the new framework provides more accurate and robust classification result as compared to commonly used likelihood methods.

EMAIL: liushuyi1028@gmail.com

Joint Models of Longitudinal Measurements and Survival Outcome with Functional Predictors for the Alzheimer's Disease Neuroimaging Initiative Study

Yue Wang*, University of North Carolina, Chapel Hill
Joseph G. Ibrahim, University of North Carolina, Chapel Hill
Hongtu Zhu, University of Texas MD Anderson Cancer Center

The use of longitudinal measurements to predict survival outcomes has a great impact in public health. In neuroimaging studies, both the longitudinal measurements and the survival outcome can be associated with imaging predictors.The aim of this paper is to develop a joint model that examines the prediction of the survival outcome by using longitudinal measurements while adjusting for a set of functional and scalar predictors. An EM estimation procedure which allows the longitudinal measurements and survival outcome to be linked through shared random effects is described. We also employ a novel approach of functional partial least square (fPLS) to obtain an more accurate estimation of the unknown slope function. Our simulation studies demonstrate the superior performance of both the EM estimators and the fPLS estimators in terms of estimation accuracy of unknown slope function. We illustrate our method in the analysis of longitudinal cognitive score, survival outcome and hippocampus data obtained from the Alzheimer's Disease Neuroimaging Initiative dataset.

EMAIL: taryue@gmail.com

Powerful, Fast, and Robust Family-Wise Error Control for Neuroimaging

Simon N. Vandekar*, University of Pennsylvania
Adon Rosen, University of Pennsylvania
Rastko Ciric, University of Pennsylvania
Theodore D. Satterthwaite, 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

Recently, several neuroimaging studies have demonstrated that multiple testing procedures (MTPs) commonly used in neuroimaging yield incorrect false positive rates. Methods that rely on Gaussian random field theory have inflated family-wise error rates (FWERs) due to the reliance on assumptions that are invalid in neuroimaging. While classical MTPs guarantee a conservative FWER they are underpowered as they ignore the covariance structure of the test statistics. We introduce a parametric bootstrap joint (PBJ) testing procedure that leverages the covariance structure of the test statistics. We use simulations to compare the PBJ procedure to nonparametric bootstrap and permutation joint testing procedures. To generate realistic data in each simulation we draw a subsample of cerebral blood flow imaging (p=109,748) and region-wise (p=112) data from the Philadelphia Neurodevelopmental Cohort (n=1,601). The PBJ procedure maintains the FWER at the nominal level and has superior power to the nonparametric procedures. Remarkably, the PBJ procedure is 150 times faster than the nonparametric bootstrap reducing computing time from 10.5 minutes to approximately 4 seconds (for n=120, p=112).

EMAIL: simonv@mail.med.upenn.edu

Nonlinear Model with Random Inflection Points for Modeling Neurodegenerative Disease Progression Using Dynamic Markers

Yuanjia Wang, Columbia University and New York State Psychiatric Institute
Ming Sun*, Columbia University

Due to a lack of a gold standard objective marker, the current practice for diagnosing a neurological disorder is mostly based on clinical symptoms, which may occur in the late stage of disease. Clinical diagnosis is also subject to high variance due to between- and within-subject variability of patient symptomatology and between-clinician variability. Under the assumption that certain markers reflect pathological processes, we propose a model that jointly estimates the changing trajectories of markers or subtle clinical signs in the same domain using subject-specific predictors. The model scales different markers into comparable progression curves with a temporal order and build the relationship between the progression of markers with underlying disease mechanism. It also assesses how subject-specific features affect the trajectories of a marker and the understanding offers guidelines for possible personalized preventive therapeutics. We applied our model to markers in cognitive, imaging and motor domains of

Huntington's disease (HD), and explained how our result can be used to predict the onset of the disease.

EMAIL: ms4799@cumc.columbia.edu

› Exploratory Analysis of High Dimensional Time Series with Applications to Multichannel Electroencephalograms

Yuxiao Wang*#, University of California, Irvine
Chee-Ming Ting, Universiti Teknologi, Malaysia
Hernando Ombao, University of California, Irvine

In this paper, we address the the major hurdle of high dimensionality in EEG analysis by extracting the optimal lower dimensional representations. Using our approach, connectivity between regions in a high-dimensional brain network is characterized through the connectivity between region-specific factors. The proposed approach is motivated by our observation that EEGs from channels within each region exhibit a high degree of multicollinearity and synchrony. We consider the general approach for deriving summary factors which are solutions to the criterion of reconstruction error. In this work, we focus on two special cases of linear auto encoder and decoder. This exploratory analysis is the starting point to the multi-scale factor analysis model for studying the brain connectivity. We performed evaluations on the two approaches via simulations under different conditions. The simulation results provide insights on the performance and application scope of the methods. We also performed exploratory analysis of EEG recorded over several epochs during resting state. Finally, we implemented these exploratory methods in a Matlab toolbox XHiDiTS available from https://goo.gl/uXc8ei .

EMAIL: yxwang87@gmail.com

124. Multivariate Survival

› Statistical Methods for Multivariate Failure Time Data Analysis

Shanshan Zhao*, National Institute of Environmental Health Sciences, National Institutes of Health
Ross L. Prentice, Fred Hutchinson Cancer Research Center

In many epidemiology studies, participants are followed for more than one health outcome. For example, in the NIEHS Sister Study, we have information about participants' related diseases (e.g., breast cancer and ovarian cancer), competing diseases (e.g., breast cancer and death), and recurrent diseases (e.g., multiple breast cancer onsets). Although univariate survival data analysis has been well developed, methods for multivariate survival data analysis are limited. Most studies analyzed these outcomes separately, or adjusted status of other diseases as covariates in the risk model of the main disease of interest. We propose to model the joint risks of multiple diseases through marginal Cox proportional hazards models and a Cox-type model for the cross-ratio. Regression parameters can be estimated by maximizing the profile likelihood. The proposed method is applied to the Women's Health Initiative's hormone therapy trial data, and it revealed a positive dependency between coronary heart disease and stroke, and also the dependency is stronger for women under hormone therapy.

EMAIL: shanshan.zhao@nih.gov

› Semiparametric Temporal Process Regression of Survival-Out-of-Hospital

Tianyu Zhan*, University of Michigan
Douglas E. Schaubel, University of Michigan

The recurrent/terminal event data structure has undergone considerable methodological development in the last 15 years. An example of the data structure that has arisen with increasing frequency involves the recurrent event being hospitalization and the terminal event being death. We consider the response Survival-Out-of-Hospital, defined as a temporal process (indicator function) taking the value 1 when the sub-

ject is currently alive and not hospitalized, and 0 otherwise. Based on estimating equations, we propose a semiparametric regression model on the joint probability of being alive and out-of-hospital. Our regression parameters estimator does not require estimating the baseline probability process over time, and the completely unspecified baseline probability process could be estimated at any time point. We demonstrate that the regression parameter estimator is asymptotically normal, and that the baseline probability function estimator converges to a Gaussian process. Simulation studies are also performed. The proposed methods are applied to the Dialysis Outcomes and Practice Patterns Study (DOPPS) Phase 5, an international study of end-stage renal disease.

EMAIL: tianyuzh@umich.edu

› Adaptive Group Bridge for Marginal Cox Proportional Hazards Models with Multiple Diseases

Natasha A. Sahr*, Medical College of Wisconsin
Soyoung Kim, Medical College of Wisconsin
Kwang Woo Ahn, Medical College of Wisconsin

Variable selection methods in linear regression such as lasso, SCAD, and group lasso have been applied to the univariate Cox proportional hazards (PH) model; however, variable selection in multivariate failure time regression analysis is a challenging and relatively unexplored research area. In this context, we propose an adaptive group bridge penalty to select variables for marginal Cox PH models with multivariate failure time data. The proposed method not only selects group variables, but also individual variables within a group. The adaptive group bridge method for marginal Cox PH models with multivariate failure time data was compared to the group bridge penalty method, and backwards, forwards, and stepwise selection. The simulation studies show that the adaptive group bridge method has superior performance compared to the other methods in terms of variable selection accuracy.

EMAIL: nsahr@mcw.edu

› Improving the Efficiency of the Proportional Rates Model

Sai H. Dharmarajan*, University of Michigan
Douglas E. Schaubel, University of Michigan

In many biomedical studies, the outcome of interest is recurrent in nature. Examples include hospitalizations, infections, and treatment failures. It is often of interest to estimate the effect of covariates on the recurrent event process. Lin et al. (2000) proposed semiparametric models for mean and rate functions of the recurrent event counting process. These models are versatile in the sense that they permit arbitrary dependence structures among events within-subject. We propose weighting methods to improve the efficiency of parameter estimates for the proportional rates model. The intention is to weight subjects inversely with respect to within-subject variance, with the weights being constant over follow-up time. Through simulation studies, we demonstrate that substantial gains in efficiency can be achieved. The methods are applied to end-stage renal disease patients receiving hemodialysis.

EMAIL: shdharma@umich.edu

› Soft-Thresholding Operator for Modeling Sparse Time Varying Effects in Survival Analysis

Yuan Yang*, University of Michigan
Jian Kang, University of Michigan
Yi Li, University of Michigan

Precision medicine calls for the development of new methods that can identify important biomarkers and model their dynamic effects on patients' survival experiences, as existing methods are neither flexible for sparsity modeling nor computationally scalable for big data analysis. We propose a new approach to estimating sparse time-varying effects of high dimensional predictors within the Cox regression framework. As opposed to the commonly used regularization methods, we propose a new soft-thresholding operator in the space of smooth functions and use it to construct sparse and piece-wise smooth time-varying coefficients. This leads to a more interpretable model with a straightforward inference procedure. We develop an efficient algorithm for inference in the target functional space and obtain

the confidence bands. We show that the proposed method enjoys good theoretical properties. The method is further illustrated and evaluated via extensive simulation studies and a data analysis of a kidney epidemiology study.

EMAIL: yuanyang@umich.edu

Copula-Based Score Test for Large-Scale Bivariate Time-to-Event Data, with an Application to a Genetic Study of AMD Progression

Yi Liu*, University of Pittsburgh
Wei Chen, University of Pittsburgh
Ying Ding, University of Pittsburgh

Motivated by a genome-wide association study (GWAS) to discover genetic risks on Age-ralated Macular degeneration (AMD) progression. We developed a copula-based score test to accommodate bivariate time-to-event data. Specifically, we consider both parametric and weakly parametric piecewise constant marginal distributions under proportional hazard assumption. Proposed method is evaluated through simulations studies under both correct model specification and misspecified scenarios when copula function or marginal distribution is wrongly assumed. Type-I error control and power performance are checked. We apply our method on Age-related Eye Disease Study (AREDS) data. Comparing it with other existing approaches. In conclusion, our proposed method is fast, robust and stable, which is ideal to apply on GWAS.

EMAIL: yiliu8927@gmail.com