Sunday, March 19, 2023 | 8:00-11:00 PM

1. POSTERS: BAYESIAN METHODS

1a. Variance as a Predictor of Health Outcomes

Irena Chen, University of Michigan

Longitudinal biomarker data and cross-sectional outcomes are routinely collected in modern epidemiology studies, with the goal of informing tailored intervention decisions. Hormones such as estradiol (E2) and follicle-stimulating hormone (FSH) may predict changes in women's health during the midlife. Most existing methods focus on predictors from mean trajectories. However, subject-level marker variability may also provide critical information about disease risk and health outcomes. Current statistical models do not investigate such relationships with valid uncertainty quantification. We propose a Bayesian joint model that estimates subject-level means, variances, and co-variances of multiple longitudinal biomarkers to predict a health outcome. The proposed method provides less biased and more efficient estimates, relative to other approaches that either ignore subject-level variability differences or perform two-stage estimation where estimated marker variances are treated as observed. Analyses of women's health data reveal that, for the first time, larger variability of E2 is associated with higher fat mass change across the menopausal transition.

1b. Transfer Learning with Uncertainty Quantification: Random Effect Calibration of Source to Target (RECaST)

Jimmy Hickey, North Carolina State University

Transfer learning uses a data model, trained to make predictions or inferences on data from one population, to make reliable predictions or inferences on data from another population. While the development of transfer learning methodology is a highly active area of research, most approaches focus on fine-tuning pre-trained neural network models that fail to provide crucial uncertainty quantification. We develop a statistical foundation for model predictions based on transfer learning; we mathematically and empirically demonstrate the validity of our approach in simple settings, and numerically illustrate the method's robustness to asymptotic approximations in more complex settings. Whereas many existing techniques are built on particular source models, our method is agnostic to the choice of source model. Our method also provides uncertainty quantification for predictions, which is mostly absent in the literature. We examine our method's performance in a simulation study and in an application to real data from the eICU Collaborative Research Database. In all cases, our approach highlights the flexibility to use different source models.

1c. Tree-Regularized Bayesian Latent Class Analysis: Subtyping Dietary Patterns Among Hispanics and Latinos

Mengbing Li, University of Michigan

Latent class models (LCMs) are widely used in nutritional epidemiology to identify subpopulations with distinct dietary patterns across various food items to assess health outcomes. The main goal is to accurately estimate the class profiles (dietary patterns) and subject-level class assignments. However, the estimated profiles often show varying degrees of similarities by pre-specified food groups (e.g., meat and grain). Agnostic to such similarities, classical LCMs often produce unstable class profile estimates, resulting in less accurate class assignments. We propose a tree-regularized Bayesian LCM that borrows information between classes guided by an unknown tree, with varying extents of shrinkage across food groups. As a key ingredient of our model, the Dirichlet diffusion tree (DDT) process specifies a prior for the unknown tree over classes. We derive a Metropolis-within-Gibbs algorithm for posterior inference. Simulation studies show that the proposed model leverages learned between-class similarity to improve model estimation relative to classical LCMs. Analyses of a dietary recall dataset among Hispanics illustrate the model's utility in dietary pattern discovery.

1d. Inference of Health Outcomes Among Patients with HIV During COVID-19 Pandemic: Using MRP Model to Improve Survey Representativeness

Amy Pitts, Columbia University, Mailman School of Public Health

The Community Health Advisory & Information Network (CHAIN) is an ongoing prospective cohort study of health outcomes among persons living with HIV in New York City. Analysis of viral load suppression, physical/mental wellness scales can give insight into the cohorts overall functioning and quality of life, as shaped by social determinants of health. However, statistical analysis of the most recent wave is challenged by missing data and potential selection bias due to implementations of sampling design and disruption in collection of data due to the COVID-19 pandemic. To improve representativeness of the recent wave, we fit a Bayesian multilevel regression with poststratification (MRP) model to correct the differences between sample and population using known population distributions obtained from the NYC HIV/AID annual surveillance statistics 2020. We considered a multilevel logistic (linear) regression for viral load suppression (physical/mental wellness scales). We showed that the sample has demographic features that are underrepresented compared to the population of interest. This deviation biases overall sample results and MRP helps improve the survey inference.

1e. Joint Bayesian Additive Regression Tree Model for Flexible Prediction from Genomic Data

Licai Huang, The University of Texas MD Anderson Cancer Center

Many diseases are heterogeneous, and different subgroups may have diverse outcome-driven biological processes. Integrative analysis using the entire dataset may miss the subtype-specific biological processes. However, analyzing the data within each subgroup may lose considerable power to identify the shared mechanisms. To address this limitation, we propose a hierarchical Bayesian model that encourages the common selection of important variables across subgroups but allows a nonidentical set of variables for each subgroup. Specifically, we use Bayesian Additive Regression Tree (BART) to characterize the key mechanism underlying the response for each subgroup. BART is a general and flexible model, allowing nonlinear effects and interactions, which may be more suitable for complex distributions of genomic data. To connect the subgroups, we impose a Markov Random Field prior on the splitting rules of BART to simultaneously select variables by borrowing strength across subgroups. Our methylation-expression association analysis results show that our model attains higher accuracy in variable selection and outcome prediction than models applied separately to each subgroup.

1f. Bayesian Goodness-of-Fit Test for Meta-Analysis of Rare Binary Events

Ming Zhang, Southern Methodist University

Random-effects (RE) meta-analysis is a crucial approach for combining results from multiple independent studies that exhibit heterogeneity. Recently, two frequentist goodness-of-fit (GOF) tests were proposed to assess the adequacy of RE model fit. However, they tend to perform poorly when encountering rare binary events. Under a general binomial-normal framework, we propose a Bayesian GOF test for meta-analysis of rare events. Our method is based on pivotal quantities that plays an important role in Bayesian model assessment and adopts the Cauchy combination idea, newly proposed in a 2019 JASA paper, to combine dependent p-values computed using posterior samples from Markov Chain Monte Carlo. The advantages of our method include straightforward conception and interpretation, incorporation of all data including double zeros without artificial correction, well-controlled Type I error, and generally higher power in detecting model misfits when compared to previous GOF methods. We illustrate the proposed method via simulation and two real data applications.

1g. A Two-Stage Bayesian Model for Assessing the Geography of Racialized Economic Segregation and Premature Mortality Across US Counties

Yang Xu, Drexel University

Racialized economic segregation, a key metric that simultaneously accounts for spatial and social polarization, has been linked to adverse health outcomes, including morbidity and mortality. Due to the spatial nature of this metric, the association between health outcomes and racialized economic segregation could also change with space. Statistical methods for measuring racialized economic segregation and health outcomes relationship are not well-developed and are usually studied at the individual level. In this paper we propose a two-stage Bayesian statistical framework that provides a broad, flexible approach to studying the spatial varying association between premature mortality and racialized economic segregation, while accounting for neighborhood-level latent health determinants in the US counties. We apply our method by using the County Health Rankings data (2020) and Public Health Geocoding Project Monograph.

WITHDRAWN 1h. Consistent Bayesian Variable Selection in High-Dimensional Hierarchical Regression

Srijata Samanta, MD Anderson Cancer Center

In several high-dimensional regression problems there is often an inherent partial ordering which stipulates that a "higher priority" predictor should be included before a "lower priority" predictor. The most common examples are time series applications where many predictors are time lags of the same variable. Penalized methods which use some nested group lasso penalties are commonly used but these can get unwieldy in high-dimensional settings. A Bayesian approach with a more compact penalty/shrinkage structure was recently developed. This method can only handle restricted chain-based hierarchy and corresponding posterior consistency results are not provided. In this work we develop novel Bayesian methodology which significantly generalizes the hierarchy structure among the predictors. We also establish high dimensional posterior model selection and estimation consistency under regularity assumptions. Furthermore, we demonstrate the statistical efficacy of the proposed approach through simulation studies and apply the same on real datasets.

1i. A Bayesian Spatial Scan Statistic for Normal Data

Laasya Velamakanni, University of South Carolina

Scan statistics are useful methods for detecting spatial clustering. While they were initially developed to detect regions with an excess of binomial or Poisson events, spatial scan statistics have been extended to detect hotspots in other types of data, including continuous and time-to-event data. They have many applications in different fields such as epidemiology (e.g. detecting disease outbreaks) and sociology (e.g. detecting crime hotpots). Spatial scan statistics identify a 'most likely cluster' and then use a likelihood ratio test to determine if this cluster is statistically significant. Spatial scan statistics have been extended to the Bayesian paradigm for different types of data such as zero-inflated count data and multivariate count data. In this work, we develop a Bayesian spatial scan statistic for normal data. We conduct a simulation study to evaluate the performance of our method under varying sample sizes, cluster sizes, and observation means. We examine the number of times we reject the null hypothesis using the Bayes factor as well as the overall sensitivity and positive predictive value.

1j. A Novel Bayesian model for Assessing Intratumor Heterogeneity of Tumor Infiltrating Leukocytes with Multi-Region Gene Expression Sequencing

Peng Yang, Rice University; The University of Texas at MD Anderson Cancer Center

Intratumor heterogeneity (ITH) of tumor-infiltrated leukocytes (TILs) is an important phenomenon of cancer biology with potentially profound clinical impacts. Multi-region gene expression sequencing data provide a promising opportunity that allows for explorations of TILs and their ITH for each subject. Although several existing methods are available to infer the proportions of TILs, considerable methodological gaps exist for evaluating ITH of TILs with multi-region gene expression data. Here, we develop ICeITH, immune cell estimation reveals ITH, a Bayesian hierarchical model that borrows cell type profiles as prior knowledge to decompose mixed bulk data while accounting for the within-subject correlations among tumor samples. ICeITH quantifies ITH by the variability of targeted cellular compositions. Through extensive simulation studies, we demonstrate that ICeITH is more accurate in measuring relative cellular abundance and evaluating ITH compared with existing methods. We also assess the ability of ICeITH to stratify patients by their ITH score and associate the estimations with the survival outcomes in two multi-region gene expression datasets from lung cancer studies.

1k. A Precision Mixture Risk Model to Identify Adverse Drug Events in Patient Subpopulations

Yao Chen, Indiana University

Despite pharmacovigilance studies demonstrate early successes on detecting adverse drug event (ADE) signals, additional knowledge of ADE risks in patient subpopulations is warranted for improving ADE prevention. Recently, the case crossover design (CCD) has been implemented for ADE detection, while controlling both observed and short-term-fixed unobserved confounding effect. In this manuscript, we propose an empirical Bayes mixture model to identify ADE signals from patient subpopulations under CCD. The proposed model mines ADE signals by estimating the posterior probabilities of null hypothesis under the empirical Bayes framework, which can be used to control false discovery rate (FDR) in high-throughput ADE mining. We illustrated the utilization of the proposed model and identified drugs associated with increased risks of ADEs only in patient subpopulations. Additionally, our simulation study demonstrated that the proposed model could control FDR at a desired level and had decent performance metrics on identifying true ADE signals. In conclusion, the proposed approach can identify ADE signals in patient subpopulations, while controlling both FDR and confounding effect.

2. POSTERS: CAUSAL INFERENCE

2a. Semi-Parametric Efficient Integrative Estimator Borrowing Historical Controls with Penalized Bias

Chenyin Gao, North Carolina State University

In recent years, real-world historical controls (HCs) have grown in popularity as a means of supplementing underpowered randomized control trials (RCTs), particularly in rare diseases or cases in which randomization is unethical or impractical. As they are not always comparable to the RCT group, such HCs may heavily bias the final decision-making if used without further scrutiny. Our paper proposes a data-driven integrated framework capable of adapting to unknown biases by linking HCs to concurrent controls. The adaptive nature is achieved by dynamically sorting out a set of comparable HCs via bias penalization. Our proposed method can simultaneously achieve (a) the semi-parametric efficiency bound when the HCs are comparable and (b) a selective dynamic borrowing feature that mitigates the impact of the existence of incomparable HCs by virtue of the semi-parametric efficiency theory and selection procedure. We establish the estimation consistency and provide a post-selection inferential technique that can produce confidence intervals with satisfying finite-sample coverage properties under extensive real data-driven simulations across various bias-generating concerns.

2b. The Probability of Causation Versus the Average Treatment Effect: Evaluating Respiratory Health and Indoor Air Pollution

Dane Isenberg, University of Pennsylvania Biostatistics (GGEB)

Public health researchers need to know whether they can attribute a negative outcome to a harmful exposure so they can make more informed policy decisions and advise on effective allocation of resources. We explore the probability of causation (PC), which addresses the question “would an exposed individual who experienced the outcome not have experienced the outcome had they not been exposed?”, in contrast to the average treatment effect (ATE). We use simulated datasets to compare the interpretation of the PC and the ATE under various real-world scenarios. We provide side-by-side analyses of the PC and the ATE using data from a cluster-randomized experiment conducted in rural India, which investigates the impact of traditional solid fuel burning stoves versus more modern stoves on respiratory health. In these analyses, we use influence-based function estimators, which allow for nonparametric estimation. We also explore the heterogeneity of PC estimates across a high-dimensional covariate space. Our comparison suggests that researchers focused on the ATE may arrive at misguided policy recommendations for circumstances that in fact require estimation of the PC.

2c. Causal Inference on Irregular Longitudinal Data

Grace Tompkins, University of Waterloo

In observational studies, we often encounter irregular longitudinal data where the number and timings of observations vary across individuals. Although there exist methods that can accommodate irregular longitudinal data, issues arise when data is collected at observation times that are driven by the longitudinal outcome. In this setting, ignoring the observation process can lead to biased estimates of the outcome model parameters. This issue can be coupled with other sources of bias, such as failing to account for non-randomized treatment assignments. Individually, there are weighting methods that can account for only one, but not both, of these sources of bias. As such, we present a new multiplicative weighting method that simultaneously accounts for both sources of bias and is simple to implement. We show that under certain conditions, this method can produce unbiased estimates of model parameters when assumptions of ignorability are violated in the observation and treatment assignment processes. We demonstrate the proposed methodology on intensive care unit data to estimate the effect of transthoracic echocardiography on nursing sentiment for patients with severe sepsis.

2d. Using Case Description Information to Reduce Sensitivity to Bias for the Attributable Fraction Among the Exposed

Kan Chen, University of Pennsylvania

The attributable fraction among the exposed (AF), also known as the attributable risk or excess fraction among the exposed, is the proportion of disease cases among the exposed that could be avoided by eliminating the exposure. Understanding the AF for different exposures helps guide public health interventions. The conventional approach to inference for the AF assumes no unmeasured confounding and could be sensitive to hidden bias from unobserved covariates. In this paper, we propose a new approach to reduce sensitivity to hidden bias for conducting statistical inference on the AF by leveraging case description information. The proposed methodology is illustrated by re-examining alcohol consumption and the risk of postmenopausal invasive breast cancer using case description information on the subtype of cancer (hormone-sensitive or insensitive) using data from the Women's Health Initiative (WHI) Observational Study (OS).

2e. Estimating Heterogeneous Causal Effects Under Bipartite Network Interference

Kevin Chen, Harvard T.H. Chan School of Public Health

Air pollution regulatory policy is largely focused on intervening on large emissions generators, while resulting health impacts are measured in the populations exposed to emissions from these sources. Moreover, the nature of air pollution transport as well as secondary pollutant formation introduces interference. Such a scenario has been described as that of bipartite network interference. Under this setting, we propose causal estimators for direct and spillover individual treatment effects via augmented inverse propensity weighting (AIPW), stabilized IPW (SIPW), and G-computation methods. Due to the complex nature of bipartite interference, to our knowledge, previous literature has not simulated data in this setting. Thus, we propose and implement a novel empirical Monte Carlo simulation scenario through which we evaluate the performance of our proposed estimators. Additionally, we apply our estimators to a study of the impacts of power plant scrubber installation on heart disease hospitalizations. Our approach provides a novel way to estimate heterogeneous causal effects through careful consideration of complex treatment and interference structures.

2f. The Correlation Between Climate Change's Effect on Precipitation and the Increase in Acute Gastrointestinal Infections

Ramanjit Sahi, Austin Peay State University

Climate change has been transforming the world in a multitude of ways. It has been presenting the life found on Earth with countless challenges. One such challenge is the alteration of precipitation patterns that has been observed in many areas. In particular, the increase in precipitation in certain areas has had significant consequences on human health. This increase in precipitation leads to changes in water source's microbial makeup via increased turbidity and surface water runoff. Higher turbidity and surface water runoff have a higher chance of contaminated with disease causing particles, animal manure etc. These changes can involve an increase in the concentrations of infectious protozoans such as giardiasis, cryptosporidiosis, cyclosporiasis, isoporiasis that cause acute gastrointestinal infections. To better understand this threat to human health, an analysis of historical data on weekly precipitation and weekly cases of giardiasis, cryptosporidiosis and cyclosporiasis was performed. It was observed that there is a significant correlation between increased precipitation and acute gastrointestinal Infections.

2g. Estimation of Conditional Average Treatment Effects for Time-to-Event Outcomes with Competing Risks

Runjia Li, University of Pittsburgh

Numerous statistical theory and methods have been proposed for estimating the causal treatment effects in observational studies. The majority of approaches can be categorized into outcome-based modeling, treatment-based modeling, and modeling for both outcome and treatment with a doubly robust feature. Currently, most of the methods with doubly robust feature do not address treatment-effect heterogeneity, specifically, estimation of personalized treatment effects in time-to-event outcomes with competing risks. In this paper, we developed a framework for estimating conditional causal average treatment effects defined as the risk difference of cumulative incidence functions given a subject's characteristics. Our method integrates targeted maximum likelihood estimation with meta-algorithms, incorporates various machine learning methods for outcome modeling and propensity score modeling, and assesses variable importance. In extensive simulation studies, our method outperformed others, even in scenarios where the outcome model was mis-specified. Application of our method is illustrated in a study of treatment effects for sepsis patients admitted to intensive care units.

2h. Benchmark Dose Contours for Bivariate Exposures via Generalized Additive Models

Tugba Akkaya Hocagil, University of Waterloo

Benchmark dose (BMD) analysis is a statistical procedure for estimating an exposure level that is associated with a specified increased risk of an adverse health outcome. We consider the challenge of BMD analysis in the context of a study aiming to relate patterns of drinking behavior in expectant mothers (e.g., proportion of days spent drinking, and the amount of alcohol consumed/drinking day) to childhood cognition. We propose a flexible framework for BMD analysis that allows a more nuanced assessment of risks associated with multi-dimensional exposure variables. The method entails fitting a generalized additive model (GAM) for the effect of the exposures on cognition while adjusting for potential confounders via suitable propensity scores to flexibly estimate the dose-response surface. From this model we obtain a benchmark dose contour that relates the two continuous exposure variables to an outcome. We illustrate our method using data assembled from six U.S. cohort studies that measured maternal reports of alcohol use during pregnancy, and measurements of cognitive function in their offspring.

2i. Propensity Score Analysis with Local Balance

Yan Li, The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences

Propensity score (PS) analysis methods rely on a correctly specified PS model. Estimation of the average treatment effect (ATE) may be biased when the model is misspecified. Nonparametric models for treatment assignment alleviate this issue, but they may not guarantee covariate balance. Methods forcing covariate mean balance between the treatment groups, termed global balance in this paper, may lead to bias in ATE estimation. Their estimated PSs only ensure global balance but not the balancing property (the conditional independence between treatment assignment and covariates given the PS). The balancing property implies not only global balance but also local balance --- the mean balance of covariates in propensity score stratified sub-populations. Local balance implies global balance, but the reverse is false. We propose the propensity score with local balance (PSLB) methodology, which incorporates nonparametric propensity score models and optimizes local balance. Extensive numerical studies showed that the proposed method can substantially outperform existing methods. The proposed method is implemented in the R package PSLB.

2j. Accounting for Time-Dependent Confounding in the Million Person Study of Low-Dose Radiation Effects

Yeji Ko, Vanderbilt University

The Million Person Study (MPS) consists of several cohorts of workers and veterans in the U.S. with the objective of

quantifying the health effects of chronic low-dose radiation exposure. Due to its observational design, MPS analyses can be sensitive to measured and unmeasured confounding. We sought to determine optimal statistical approaches to account for time-dependent confounding. First, in simulation studies based on the MPS data, we compared the performance of alternative methods to estimate the association of a time-dependent exposure with a time-to-event outcome in the presence of time-depending confounding. We fit standard Cox models with adjustment for the confounder and Cox marginal structural models based on propensity scores. Next, we applied these methods to workers of Los Alamos National Laboratory to estimate the association of time-dependent cumulative radiation dose with all-cause mortality, with consideration of time-depending confounding by duration of employment. Radiation epidemiology studies, particularly those focused on radiation effects at low doses, should consider specialized methods to account for time-dependent confounding.

2k. Efficient Generalization and Transportation

Zhenghao Zeng*, Carnegie Mellon University

When estimating causal effects, it is important to determine how useful a given randomized trial/observational study is to inform a practical question for a specific target population. If samples collected cannot represent the target population well, the average treatment effect (ATE) obtained from trials/study may not generalize to the target population directly. To tackle this problem, we propose new methods to generalize or transport the ATE from trial/observational study to the target population in the case where the population underlying the trial/observational study and the target population have different sets of covariates. When the ATE in the target population is identified, we propose doubly robust estimators based on efficiency theory and establish its asymptotic normality under further conditions. Simulation study shows the advantage of the proposed doubly robust estimator over the plug-in estimator. The proposed methods are applied in transporting causal effects of dietary intake on adverse pregnancy outcomes from an observational study to the whole U.S. female population as an illustration.

2l. Asymmetric Predictability in Causal Discovery: An Information Theoretic Approach

Soumik Purkayastha, University of Michigan

Causal discovery in observational studies pose great challenges in scientific research where randomized trials or intervention-based studies are not feasible. This project is motivated by a cohort study which investigates if environment-triggered changes in epigenetic markers influence cardiovascular function or if the converse is true. We develop an information theoretic causal discovery framework of predictive asymmetry - whether X is a stronger predictor than Y or vice-versa. Our framework introduces a new statistic called the Directed Mutual Information (DMI). The DMI detects complex association in bivariate (X, Y) and quantifies predictive asymmetry, aiding our notion of causality. Our framework relies on density estimation, that is done by a Fourier transformation-based approach. Our method is many magnitudes faster than the classical bandwidth-based density estimation method, while maintaining comparable error performance, making our method scalable. We establish key large-sample properties of our framework by developing a new data-splitting inference technique and evaluate its performance through simulation studies and a real data example in the motivating study.

2m. Nonparametric Estimation of Conditional Incremental Effects

Alec McClean, Carnegie Mellon University

We analyze conditional effects based on incremental propensity score interventions, which consider the effect of counterfactually multiplying the odds of treatment by some factor. A benefit of these effects is that they are still identifiable even when some subjects deterministically receive treatment. We develop projection estimators and flexible nonparametric estimators and derive model-agnostic error guarantees showing they satisfy a form of double robustness. We also propose a summary of treatment effect heterogeneity and derive a nonparametric estimator satisfying a form of double robustness. We demonstrate our estimators on real data by analyzing the effect of ICU admission on mortality.

3. POSTERS: CLINICAL TRIALS AND BIOPHARMACEUTICAL RESEARCH METHODS

3a. Statistical and Machine Learning Methods for Adaptive Radiotherapy Treatment Scheduling

MaryLena Bleile, Southern Methodist University

Pre-clinical cancer studies have revealed substantial room for improvement in the radiation fractionation schemes applied in radioimmunotherapy regimens. This work aims to develop and evaluate a generalized, adaptive method to identify the optimal radiation regimen for use with immunotherapy. We accomplish this using a method that can support classical statistical models as well as modern machine learning. We evaluate and compare three versions of our method in a simulation experiment that models an adaptive in vivo tumor xenograft study. We observe that the predictive system characterized by a linear spline model is most efficient, robust, and conservative, i.e. safest to apply in a real tumor xenograft experiment.

3b. Exact Sequential Single-Arm Trial Design with Curtailment for Binary Endpoint

Tasuku Inao, Hokkaido University

For trials for rare cancer, the single-arm design with binary response is preferred incorporating sequential monitoring. It tends to be hard to conceal the number of responders completely, and it may be appropriate to present the prespecified thresholds and to monitor each patient to determine if the trial is a success or not. In this paper, we propose the statistical monitoring method with the threshold of responders for efficacy stopping fixed over the number of participants. To maintain the error rate, it is exactly calculated using the negative binomial distribution. The bias-adjusted point estimator is also proposed. The simulation experiments showed that the maximum sample size of our exact sequential design was less than an exact fixed design, and the average sample number of our design was as small as conventional Simon's two-stage design or spending function approach. Also, we observed a small bias in proposed estimator and the convenient methods of confidence interval.

3c. Building a Foundation for More Flexible A/B Testing: Applications of Interim Monitoring to Large Scale Data

Wenru Zhou, University of Colorado, Anschutz campus

The use of error spending functions and stopping rules has become a powerful tool in interim analysis. The implementation of interim analysis is broadly desired not only in traditional clinical trials, but also in A/B tests. Although many papers have summarized error spending approaches, a comprehensive review is needed, targeting large-scale data, to help people in industry find their optimal boundary easily. In this paper, we summarized sixteen existing boundaries including fifteen boundaries that consist of five error spending functions that allow early termination for futility, difference, or both, as well as a fixed sample size design is included as a boundary. The simulation is based on a practical A/B testing problem comparing two independent proportions. Sample sizes change from approximately 500 to 250,000 per arm to handle different sample sizes in practice. The choices of optimal boundaries are summarized using a loss function that incorporates different weights of expected sample size under null, alternative, and maximum sample size. The results based on adequate power design, under-powered, and over-powered design are presented in tables and figures.

3d. Modified Isotonic Regression Based Phase I/II Clinical Trial Designs Identifying Optimal Biological Dose

Yingjie Qiu, Indiana University

Conventional phase I/II clinical trial designs often use complex parametric models to characterize the dose-response relationships. However, the parametric models are hard to justify in practice, and the misspecification of parametric models can lead to substantially undesirable performances in phase I/II trials. It is difficult for physicians to clinically interpret the parameters of these complex models, and such significant learning costs impede the translation of novel statistical designs into practical trial implementation. To solve these issues, we propose a transparent and efficient phase I/II clinical trial design, referred to as the modified isotonic regression-based design (mISO), to identify the optimal biological doses for molecularly targeted agents and immunotherapy. The mISO design makes no parametric model assumptions and yields desirable performances under any clinically meaningful dose-response curves. We further extend the mISO design and develop the mISO-B design to handle the delayed outcomes. Our comprehensive simulation studies show that the mISO and mISO-B designs are highly efficient and outperform many existing phase I/II clinical trial designs.

3e. Futility Assessment in Sequential Clinical Trial with Multiple Binary Endpoints by Using the Predictive Probability of Success

Weiyi Xia, Rutgers University

Predictive Probability of Success (PPOS), is a commonly used method in clinical trials to support decision-making under the Bayesian framework. PPOS is well established and discussed for single and multiple continuous endpoints. However, there is little work regarding multiple binary endpoints. Motivated by a real-life example, we propose a linear transformation to the multiple binary endpoints to obtain a univariate multinomial endpoint. With the purpose of assessing the probability of success after an interim analysis for a single-arm group sequential design, a simulation study using 2 binary endpoints was performed. The impact of correlated binary endpoints was also evaluated. The results show PPOS enables the futility monitoring decision-making under different stopping rules in the trial with 2 binary endpoints.

3f. An Extended Bayesian Semi-Mechanistic Dose-Finding Design for Phase I Oncology Trials Using Pharmacokinetic and Pharmacodynamic Information

Chao Yang, The University of Texas MD Anderson Cancer Center

We propose a model-based, semi-mechanistic dose-finding (SDF) design for phase I oncology trials that incorporates pharmaco-kinetic/dynamic (PK/PD) information when modeling the dose-toxicity relationship. This design extends a recently proposed SDF model framework by incorporating measurements for a PD biomarker relevant to the primary dose-limiting toxicity (DLT). We propose joint Bayesian modeling of the PK, PD, and DLT outcomes. Our extensive simulation study shows that on average the proposed design outperforms common phase I trial designs, including modified toxicity probability interval (mTPI) and Bayesian optimal interval (BOIN) designs, the continual reassessment method (CRM), as well as an SDF design assuming a latent PD biomarker, in terms of the percentage of correct selection of maximum tolerated dose (MTD) and average number of patients allocated at MTD, under a variety of dose-toxicity scenarios. The proposed design also yields better estimated dose-toxicity curves than CRM. A sensitivity analysis suggests the design's performance is robust to prior specification for the parameter in the link function between cumulative PD effect and DLT probability.

3g. The Effect of Subject Accrual on Futility Analyses in Randomized Clinical Trials with Time-to-Event Endpoints: A Simulation Study on Conditional Power

Corinne McGill, The Medical University of South Carolina - Charleston, SC

The longitudinal nature of most clinical trials presents the need for interim monitoring. One method to determine early stopping for futility is conditional power, defined as the probability the null hypothesis will be rejected at the trial's end based on the fraction of data available at interim analysis and assumptions about the remaining data. In the log-rank setting for time-to-event endpoints, this information fraction is a function of two event counts: observed and expected. The expected event count is calculated from hypothesized rates and estimated amounts of follow-up accumulated per subject. In studies with variable follow-up time, departures from assumed uniform subject accrual will affect accumulated follow-up, potentially impacting observed events. Hence, lagging enrollment can have a negative effect on power. Its impact on conditional power was examined under various accrual patterns in a simulation study for a two-arm superiority trial. For interim analyses conducted at calendar-based intervals, conditional power is erroneously estimated in certain scenarios when the assumption of uniform enrollment is not met, highlighting the need to account for non-uniformity.

3h. Marginal Regression on Transient State Occupation Probabilities with Clustered Multistate Process Data

Wenxian Zhou, Indiana University

Clustered multistate process data are commonly encountered in multicenter studies. A clinically important estimand with such data is the marginal probability of being in a particular transient state as a function of time. However, there is currently no method for marginal regression analysis of these probabilities with clustered multistate process data. To address this problem, we propose a weighted functional generalized estimating equations approach that does not impose Markov assumptions or assumptions regarding the structure of the within-cluster dependence and allows for informative cluster size (ICS). The asymptotic properties of the proposed estimators for the functional regression coefficients are rigorously established, and a nonparametric hypothesis testing procedure for covariate effects is proposed. Simulation studies show that the proposed method performs well and that ignoring the within-cluster dependence and the ICS leads to invalid inferences. The proposed method is used to analyze data from a multicenter clinical trial on recurrent or metastatic squamous-cell carcinoma of the head and neck with a stratified randomization design.

4. POSTERS: EPIDEMIOLOGICAL METHODS

4a. A Likelihood-Based Inference Approach for Detecting Differential Item Functioning with Application in Mild Cognitive Impairment

Zeling He, Emory University

Functional Assessment Questionnaire (FAQ) measures instrumental activities of daily living to help the diagnosis of mild cognitive impairment (MCI). Despite its usefulness, FAQ is self-reported and may be biased for different demographic groups. This bias is called differential item functioning (DIF). Existing methods only detect DIF in one group at a time or require repeated model fittings to conduct inference for multiple groups. In this article, we developed an efficient likelihood-based inference approach for detecting DIF that provided accurate estimations and performed inference for multiple demographic groups simultaneously. The simulation results showed that our proposed procedure outperforms the existing methods in estimation accuracy, with well-controlled type I error and robust power. We applied the proposed method to detect DIF of FAQ in different race and gender groups to MCI data from the National Alzheimer's Coordinating Center. Our analysis showed that all items in FAQ presented bias in race. Compared with other subgroups of the same impaired functioning level, black males were less likely to report functional impairment in FAQ.

4b. Group Specific Dynamic Models of Time Varying Exposures on A Time-to-Event Outcome

Yan Tong, Indiana University

Assessing cumulative effects of time-varying exposures poses challenges. This study provides a novel approach to modeling group-specific dynamics between cumulative time-varying exposures and a time-to-event outcome. A framework of group-specific dynamic models is introduced utilizing functional time-dependent cumulative exposures within a relevant time window. Penalized-spline time-dependent Cox models are proposed to evaluate group-specific outcome-exposure dynamics through the associations of a time-to-event outcome with functional cumulative exposures and group-by-exposure interactions. Model parameter estimation is achieved by penalized partial likelihood. Information criteria AIC, AICc, and BIC are used for best-fitting spline selection. Hypothesis testing for comparison of group-specific exposure effects is performed by Wald type tests. Extensive simulation studies are conducted and demonstrate satisfactory model performances. The proposed method is applied to the analyses of group-specific associations between antidepressant use and coronary artery disease in a depression-screening cohort using data extracted from electronic medical records.

4c. The Impact of Correlated Exposures and Missing Data on Multiple Informant Models Used to Identify Critical Exposure Windows

Jemar Bather, Harvard T.H. Chan School of Public Health

There has been interest in identifying critical windows of exposure for adverse health outcomes. Multiple informant models implemented using generalized estimating equations (MIM GEEs) have been applied to address this research question because they enable statistical comparisons of differences in associations across exposure windows. As interest rises in using MIMs, the feasibility and appropriateness of their application under settings of correlated exposures and partially missing exposure measurements requires further examination. We evaluated the impact of correlation between exposure measurements and missing exposure data on the power and differences in association estimated by the MIM GEE and an inverse probability weighted extension to account for informatively missing exposures. We showed that applying MIM GEEs maintains higher power when there is a single critical window of exposure and exposure measures are not highly correlated, but may result in low power and bias under other settings. We applied these methods to a study of pregnant women living with HIV to explore differences in association between trimester-specific viral load and infant neurodevelopment.

4d. Associations of Physical Activity, Racial Disparities, Social-Economic Status and Other Comorbidities with Nonalcoholic Fatty Liver Disease in NHANES 2003-2006

Lucia Tabacu, Old Dominion University

Nonalcoholic fatty liver disease (NAFLD) is the most common cause of chronic liver disease in the United States. NAFLD is commonly associated with metabolic conditions such as diabetes, obesity and cardiovascular disease. Since current pharmacologic treatments are not effective against NAFLD, lifestyle changes such as increased physical activity and improved diet are recommended. This talk will focus on predictive models that describe the relationships between different accelerometry-derived physical activity measures and NAFLD, accounting for racial disparities, socio-economic status and other comorbidities using the NHANES cohorts 2003-2006.

4e. Estimation of Conditional Treatment Effect and Prediction of Binary Outcomes Using the Joint Use of Propensity and Prognostic Scores

Jonggyu Baek, UMass Chan Medical School

Marginal treatment effect (MTE), including the average treatment effect on treated (ATT) and the average treatment effect (ATE), informs policy makers of a treatment’s effect at the population level. Relative to MTE, conditional treatment effect (CTE) and patient outcome prediction are more informative at the individual level. In clinical settings with shared decision making, patients and clinicians alike may want to know the effect of a treatment and resulting health outcome in the context of the patient’s specific health status and comorbidities. To estimate CTE on the individual level, we have proposed a method that employs a generalized additive model (GAM) with a tensor product smoother of estimated propensity and prognostic scores. Using this method to control for confounders and to utilize patient’s conditions, we posit that outcome prediction in the setting of observational studies will be improved. In order to apply this proposed method to practice, we examined the association between 30-day statin discontinuation and 1- and 2- year mortality among newly admitted nursing home residents in the US, 2011-2016.

4f. Quantifying Improvements in the Loss in Expectation of Life Due to Cancer in the US Using Flexible Parametric Survival Models

Theresa Devasia, National Cancer Institute

Cancer is becoming a chronic disease, largely due to improvements in treatment for multiple cancer sites. Life expectancy is a popular measure for the general population, but it has had limited use in the cancer survival literature. The loss in expectation of life (LEL) due to cancer is defined as the difference between the general population life expectancy and the life expectancy among cancer patients. To obtain life expectancy of the cancer cohort, we applied the flexible parametric survival model, which models the log cumulative excess hazard and extrapolates individual relative survival data using spline models. LEL is calculated as the difference in area underneath the general population expected survival and cancer patient overall survival curves. We selected female breast cancer, Chronic myeloid leukemia (CML), colorectal cancer (CRC), diffuse large B-cell lymphoma (DLBCL), and melanoma data between 1975 and 2018 from nine Surveillance, Epidemiology, and End Results registries. Large decreases in LEL were observed between 1990 and 2010 for female breast cancer, DLBCL, and CML. The decreases in LEL for each cancer cohort correspond to advancements in treatment.

5. POSTERS: FUNCTIONAL DATA ANALYSIS

5a. Exploring Distributional Representation of Wearable Data

Pratim Guha Niyogi, Johns Hopkins Bloomberg School of Public Health

Advancement of mobile digital health technologies (mHeath) including wearables open up exciting opportunities for better understanding links between human behaviors and human health. Wearable technology includes electronic devices that collect personal health and behavioral data such as physical activity, sleep, heart rate in real time in real life setting. Durations of walking bouts or different sleep stages are often extracted from continuous wearable streams to quantify daily activity or nighttime sleep. Motivated by Stanford Technology Analytics and Genomics in Sleep Study of 1853 individuals with various sleep disorders, we propose alternative distributional representations, apply them to quantify digital biomarkers of main stages of sleep and explore their predictive performance for discriminating subjects with different sleep disorders.

5b. Regularized Simultaneous Estimation of Changepoint and Functional Parameter in Functional Accelerometer Data Analysis

Margaret Banker, University of Michigan

Accelerometry data enables scientists to extract personal digital features useful in precision health decision making. Existing analytic methods often begin with discretizing Physical Activity (PA) counts into activity categories via fixed cutoffs; however, the cutoffs are validated under restricted settings and cannot be generalized across studies. Here, we develop a data-driven approach to overcome this bottleneck in the analysis of PA data, in which we holistically summarize an individual's PA profile using Occupation-Time Curves that describe the percentage of time spent at or above a continuum of activity levels. The resulting functional curve is informative to capture time-course individual variability of PA. We investigate functional analytics under an L0 regularization approach, which handles highly correlated micro-activity windows that serve as predictors in a scalar-on-function regression model. We develop a new one-step method that simultaneously conducts fusion via change-point detection and parameter estimation through a new L0 constraint formulation, which we evaluate via simulation experiments and a data analysis assessing the influence of PA on biological aging.

5c. Sources of Residual Autocorrelation in Multiband Task fMRI and Strategies for Effective Mitigation

Fatma Parlak, Indiana University

In task fMRI analysis, ordinary least squares (OLS) is typically used in a linear regression model to estimate task-induced activation in the brain. Since task fMRI residuals often exhibit temporal autocorrelation, it is common practice to perform prewhitening; prior to OLS to satisfy the assumption of residual independence. In this article, we first thoroughly examine the sources of residual autocorrelation in modern fast-task fMRI. We find that it varies spatially and is affected by the task, the acquisition method, modeling choices, and individual differences. Second, we evaluate the ability of different prewhitening strategies to mitigate autocorrelation and false positives. We consider two factors: the choice of autoregressive (AR) model order and the regularization of AR model coefficients, ranging from local smoothing to global averaging. We find that local regularization is much more effective than global averaging at mitigating autocorrelation. Increasing the AR model order is also helpful but to a lesser degree. To overcome the computational challenge associated with spatially variable prewhitening, we developed a fast R implementation based on parallelization and C++.

5d. Generalized Functional Linear Regression Model with Functional and Scalar Covariates Measured with Measurement Error

Yuanyuan Luan, School of Public Health, Indiana University

While extensive work has been done to correct for biases due to measurement error in scalar-valued covariates prone to errors in generalized linear regression models, limited work has been done to address biases associated with functional covariates prone to errors or the combination of scalar and functional covariates prone to errors in generalized linear regression models. In this work, we propose semiparametric and parametric approaches to correct for measurement errors associated with a mixture of functional and scalar covariates prone to errors in generalized linear regression. The developed methods are applied to investigate the influence of wearable device-based physical activity and self-reported measures of dietary intake on the probability of type 2 diabetes diagnosis. We treat the device-based measures of physical activity as error-prone functional covariates prone to complex arbitrary heteroscedastic errors. While the dietary intake is considered a scalar-valued covariate prone to error. We present simulation studies to assess the finite sample properties of our proposed methods.

6. POSTERS: HIGH DIMENSIONAL AND CLUSTERED DATA METHODS

6a. High-Dimensional Mean Vector Test for One-Sided Hypothesis

Rongrong Wang, Medical College of Georgia, Augusta University

The advancement of data acquisition technologies and computing resources have greatly facilitated the analysis of massive data sets in various fields. A unique characteristic of these data sets contains a large number of features but a small number of subjects, known as high-dimensional data. These data demand new statistical methods to enhance scientific knowledge. One of the important statistical inferences is mean vector testing. A lot of efficient statistical methods have been developed for performing the two-sided mean vector test. The one-sided high-dimensional mean vector test has received limited attention. One relevant application could be identifying significant upregulated or downregulated gene sets from the preselected gene sets. This work develops a procedure for a one-sided high-dimensional mean vector test, known as the generalized max-component test (GMCT). We study the asymptotic distribution of GMCT statistics. The GMCT is computationally efficient and robust to heteroscedasticity in the component variances. The finite sample performance of the proposed test statistic is evaluated, and it achieves competitive rates for type-I error and power.

6b. Diffusion-Enhanced Partition-Level Integrative Similarity Learning for Disease Subtyping

Yuqi Miao, Columbia university

Integrating multi-omics data to identify disease subtypes has been shown to increase the clustering accuracy and help identify novel disease subtypes. Many studies integrate subject-level similarity networks from each data type and then perform clustering on the integrated network. However, the noisy nature of the high-dimensional similarity measures and heterogeneity among dimensions of multi-omics data make integration challenging. Here we introduce a new kernel-based clustering framework, called Diffusion-Enhanced Partition-level integrative SIMiLaRity learning (DEP-SIMLR). DEP-SIMLR diffuses individual networks using local structures to reduce noise and then learns integrative partition/cluster information from each diffused network. We conducted simulation studies to examine the performance of the proposed method. DEP-SIMLR shows high cluster accuracy when different omics data types are noisy and have different dimensions. We applied the proposed and competing methods to The Cancer Genome Atlas (TCGA) program data to subtype kidney renal papillary (KIRP) cancer patients. Subtypes obtained from DEP-SIMLR have more significant differences in the survival outcome.

6c. Local Bayesian Modeling Approach for Simultaneous Estimation and Feature Extraction with Application to Sparse, High Dimensional Spatio-Temporal Data

Garrett Frady, University of Connecticut

As a result of the vast advancements in technology, we frequently come across data in high dimensions. We propose to extend the utility of the Gaussian and Diffused-gamma (GD) prior for feature extraction when dealing with sparse, high dimensional spatio-temporal data. To bypass the computational complexity, we build local binary classification models of subject-level responses at each time point using logistic regression and incorporate the temporal structure through our subject-level prediction process. The effectiveness of our method will be demonstrated through a simulation study. We will also conduct a case study with multi-subject electroencephalography (EEG) data to identify active regions of the brain and predict the risk of early-onset alcoholism. One goal of EEG analysis is to extract information from the brain in a spatiotemporal pattern and analyze the functional connectivity between different areas of the brain as a response to a certain stimulus. Selecting active regions of the brain can be viewed as a feature extraction process.

6d. Rank Intraclass Correlation for Clustered Data

Shengxin Tu, Vanderbilt University

Clustered data are common in biomedical research. Observations in the same cluster are often more similar to each other than to observations from other clusters. The intraclass correlation coefficient (ICC), first introduced by R. A. Fisher, is frequently used to measure this degree of similarity. However, the ICC is sensitive to extreme values and skewed distributions, and depends on the scale of the data. It is also not applicable to ordered categorical data. We define the rank ICC as a natural extension of Fisher's ICC to the rank scale, and describe its corresponding population parameter. The rank ICC is simply interpreted as the rank correlation between a pair of observations from the same cluster. We also extend the definition when the underlying distribution has more than two hierarchies. We describe estimation and inference procedures, show the asymptotic properties of our estimator, conduct simulations to evaluate the performance of our estimator, and illustrate our method in three real data examples with skewed data, count data, and three-level data.

6e. Bregman Divergence-Based Data Integration with Application to Polygenic Risk Score (PRS) Heterogeneity Adjustment

Qinmengge Li, University of Michigan Ann Arbor

Polygenic risk scores (PRS) have recently received much attention for genetics risk prediction. While successful for the Caucasian population, the PRS based on the minority population suffer from small sample sizes, high dimensionality and low signal-to-noise ratios, exacerbating already severe health disparities. Due to population heterogeneity, direct trans-ethnic prediction by utilizing the Caucasian model for the minority population also has limited performance. In addition, due to data privacy, the individual genotype data is not accessible for either the Caucasian population or the minority population. To address these challenges, we propose a Bregman divergence-based estimation procedure to measure and optimally balance the information from different populations. The proposed method only requires the use of encrypted summary statistics and improves the PRS performance for ethnic minority groups by incorporating additional information. We provide the asymptotic consistency and weak oracle property for the proposed method. Simulations and real data analyses also show its advantages in prediction and variable selection.

6f. Variable Selection and Prediction of C-peptide Decline in Type 1 Diabetes Using MicroRNA Data

Xuan Chen, University of Miami

Type 1 diabetes is a chronic autoimmune disease resulting in severely impaired insulin secretion. Despite the emergence of reproducible associations of circulating microRNAs (miRNAs) with type 1 diabetes, there are limited data about miRNA prediction of C-peptide decline after diagnosis. In this study, we investigated the association between miRNA and C-peptide AUC decline after diagnosis across different cohorts and sequencing platforms. A new framework of variable selection (VarPro) was applied for variable selection and compared with minimal depth and variable hunting method in random forest. Bootstrapping and cross-validation method were applied in method selection and parameter tuning. Then a random forest was built for prediction incorporated identified miRNAs and clinical characteristics. Consequently, miRNAs identified can improve prediction performance of C-peptide AUC decline. More interestingly, the miRNAs identified differ across sequencing platforms and cohort samples. This study suggests that miRNAs may be useful in predicting future C-peptide decline for advancing therapeutic discoveries for type 1 diabetes.

6g. Stagewise Majorization Minimization for Fast Regularized Learning

Boyang Tang, University of Connecticut

“Regularization + optimization” is a popular learning scheme in statistics. However, its computation could be intensive with complex loss functions. We develop a general stagewise learning strategy, as a slow-brewing process for model building, to efficiently approximate the solution paths of a sparse regularization problem with a general loss function. The main idea is to combine stagewise learning with majorization-minimization (MM), so that each incremental update is based on efficiently optimizing a surrogate loss. We present a general stagewise algorithm and perform rigorous convergence analysis, to bridge the stagewise MM algorithm and the corresponding regularized estimation problem and to reveal the trade-off between solution accuracy and computational efficiency. Practical stopping criteria and choices of step size are discussed. We use several examples, including penalized logistic regression and penalized linear mixture model to showcase the proposed approaches.

6h. Low-Rank Matrix Estimation in the Presence of Change-Points

Lei Shi*, University of California, Berkeley

Modern biomedical studies involve analyzing time-varying and high-dimensional measurements. Conventional methods have difficulty in handling such data due to the limited sample size and the non-stationary, large-scale features. In this manuscript, we study the use of low rank matrix estimation and change-point detection methods for meaningful representation and systematic analysis of matrix-variate data sequence. We consider a general trace regression model with multiple structural changes and propose a framework for simultaneous low-rank signal recovery and structural break detection. Our method incorporates nuclear norm penalized minimization into a grid search scheme that determines the potential changes along the data stream. Theoretically, we establish the non-asymptotic error bounds with a nearly-oracle rate for the matrix estimators as well as the super-consistency rate for change-point detection. As an application, we apply the joint estimation-detection framework to an air pollution study regarding inhalable particulate matter to explain the interplay between pollutants and detect changes along time to facilitate public health policy making.

6i. The Non-Overlapping Statistical Approximation to Overlapping Group Lasso

Mingyu Qi, University of Virginia

Group lasso is a commonly used regularizer in statistical learning in which parameters are eliminated from the model according to predefined groups. However, when the groups overlap, optimization of group lasso can be time-consuming because of the non-separability induced by the overlapping groups. This bottleneck has seriously limited the application of overlapping group lasso regularization in many modern problems. In this paper, we propose a separable penalty as an approximation of the overlapping group lasso penalty. Thanks to the separability, the computation of regularization based on our penalty is substantially faster than that of overlapping group lasso. We show that the penalty is the tightest separable relaxation of the overlapping group lasso norm. Moreover, we show that the estimator based on the proposed penalty is statistically equivalent to the one based on the overlapping group lasso penalty with respect to their error bounds and the rate-optimal performance. We demonstrate this efficiency and statistical equivalence in simulation examples and a classification problem of cancer tumors based on gene expression and pathways.

7. POSTERS: IMAGING AND NEUROSCIENCE

7a. Application of Closed-Form Gamma Mixture Model in mxIF Cell Gating

Jiangmei Xiong, Vanderbilt University

Multiplexed immunofluorescence (mIF) imaging is a sub-cellular resolution technology where cell expression levels are captured with multichannel images of stained tissue samples that identify up to 30 proteins. Marker gating is the process of identifying cell phenotypes based on cell marker expression values. Traditional manual gating is slow, and results are not reproducible, while recent software for phenotyping assumes a log-normal model, which is not appropriate for non-modal cell population densities. We introduce a closed-form gamma mixture model to perform cell gating, which removes the subjectivity introduced by traditional manual gating procedures and can easily incorporate biological information while speeding up the estimation process. In this work, we derived the closed-form gamma mixture model, and illustrate the analysis pipeline to perform automatic gating of segmented cell data. As the model is applied to a large number of slides and marker channels, we introduce diagnostic plots that aid in fine-tuning the results. We also evaluate our method by comparing with silver standard and another recently developed method for phenotyping mIF data.

7b. Heterogeneity Adjustment in Longitudinal Imaging Studies

Harshita Dogra, Florida State University

Due to the differences in instrumental setups, image acquisition protocols, design, and other unknown hidden factors, the heterogeneity has been commonly observed in existing large-scale imaging studies, such as Alzheimer's disease neuroimaging initiative (ADNI) study. It is even more challenging to handle the heterogeneity in longitudinal neuroimaging studies because of the unobserved effects of hidden factors on the neuroimaging data and their complex spatial-temporal structure. In this paper, we propose a longitudinal functional hybrid factor regression model that can successfully address this challenge and investigate the relationship between longitudinal imaging data and covariates of interest. We also propose both estimation and inference procedures for the hidden factors and unknown parameters. Some simulation studies and real data analysis are conducted to assess the finite sample performance of our model.

7c. Correcting Inter-Scanner Biases in High-Dimensional Neuroimaging Data via Gaussian Process

Rongqian Zhang, University of Toronto

In neuroimaging studies, combining data collected from multiple study sites has become common to increase reproducibility of scientific discoveries. However, in the process of combining, unwanted variations arise from using different scanners in data acquisitions, termed "inter scanner biases". Despite existing methods for correcting scanner effects (e.g. ComBat), most of these methods only focus on removing scanner-specific means and variances. They are limited to handle high-dimensional data that reveals a strong spatial autocorrelation, such as cortical thickness. To address these challenges, we develop a novel multivariate normalization method that models and reduces inter-scanner biases for vertex-level cortical thickness data. We evaluate and compare our method's performance to existing methods, in terms of removing scanner effects, retaining spatial auto-correlation, and statistical power, using extensive simulation studies.

7d. Leveraging Multimodal Neuroimaging Data to Identify Novel Genetic Pathways to Alzheimer's Disease

Yuan Tian, University of Toronto

Recent genome-wide association studies (GWASs) have identified multiple genetic risk factors for Alzheimer's disease (AD). However, they do not provide a comprehensive understanding of how genetically-regulated structural and functional brain pathways drive AD progression, which is critical for characterizing the genetic mechanism of AD and developing AD targeted therapeutics. We propose a three-step mediation method for analyzing causal pathways of gene-AD effects by incorporating high-dimensional and multimodal brain magnetic resonance imaging (MRI) measures as mediators in GWAS. First, to reduce high dimensionality of MRI while preserving shared structure across brain modalities, we apply BIDIFAC, a dimension reduction method that extends the idea of principal components (PCA). Next, we estimate the genetic variations of reduced-dimensional MRI features using penalized regression. Lastly, we test for the existence of intermediate causal effects between genes and AD with an adaptive association test. We apply the proposed method to UK Biobank (UKB) and International Genomics of Alzheimer's Project (IGAP) data to identify novel genetic pathways to AD.

7e. Multiple Testing Methods Using Multi-layer Networks for Detecting Disease-Associated Brain Regions and Functions in Disease-association Studies with Neuroimaging Data

Ryo Emoto, Nagoya University

In disease-association studies employing voxel-level inference using neuroimaging data, biological inference of disease-related functions is typically made for the brain regions containing the detected voxels. In this paper, we consider a model with multiple layers of latent variables as a model of disease association for the entire brain, which captures the dependent structure between voxels by assuming a network between voxels and functions. In addition, the model can incorporate the association between the functions by adding a layer of large-scale brain networks. Based on this model, we derive procedures for multiple testing and estimation of disease associations for both individual voxels and brain functions. As such, the proposed method enables quantitative evaluation on disease association for detected functions, as well as detected regions, in the context of voxel-level, disease-association analysis.

7f. Model Selection for Exposure-Mediator Interaction

Ruiyang Li, Columbia University

In mediation analysis, the exposure often influences the mediating effect, i.e., there is an interaction between exposure and mediator on the dependent variable. When the mediator is high-dimensional, it is necessary to identify non-zero mediators (M) and exposure-by-mediator (X-by-M) interactions. Although several high-dimensional mediation methods can naturally handle X-by-M interactions, the research is scarce in preserving the underlying hierarchical structure between the main effects and the interactions. To fill the knowledge gap, we develop the XMInt procedure to select M and X-by-M interactions in the high-dimensional mediators setting while preserving the hierarchical structure. Our proposed method employs a sequential regularization-based forward-selection approach to identify the hierarchically related mediators and their interactions with exposure. Our numerical experiments showed promising selection results. Further, we applied our method to ADNI morphological data and examined the role of cortical thickness and subcortical volumes on the effect of amyloid-beta accumulation on cognitive performance.

8. POSTERS: LONGITUDINAL DATA ANALYSIS

8a. Opioid Monitoring with an Example in an Alabama Rural County

Yuhui Yao, The University of Alabama

Alabama, being the state in the nation with the highest opioid prescription per-capita, suffers from the crisis of opioids as many other states in the U.S. The crisis is seen to be contributed mainly by two sources, illicitly prescribed and illegally obtained opioids. The misuse of illegal drugs has been seen to rise in opioid-related deaths, especially among African Americans. In rural Alabama, the mortality may even be worse because of poverty and health care disparities. This epidemic needs to be stopped. In order to better understand, control and reduce the epidemic of opioids, one of the effective ways is to monitor opioid-overdose-related indicators. In this work, we propose and study an advanced analytical method to improve the monitoring of the opioid-related ER visits. The methodology will be built using a monitoring scheme in Statistical Process Monitoring (SPM). In order to make the analysis more timely and actionable, the method will be applied to the daily county-based time-series count data using a Bayesian Partially Ordered LASSO model. An illustration will be provided using the data from the Walker County, which has Alabama's highest opioid prescription rate.

8b. Longitudinal Analysis of Changing Restaurant Advertising on Obesity Risk and Disparities in US Adults

Qiuyue Kong, Harvard Chan School of Public Health

Food environments are associated with obesity, which is a major risk factor for cancer and disproportionately impacts poor and racial minority populations. Restaurants are a major contributor to the food environment. We hypothesize that increased restaurant advertising may target these populations at greater risk for obesity. Using geographic and quarterly expenditure data from the top 100 US grossing restaurant chains from 2012-2016, as well as census data on US counties, we create an objective measure of local per capita restaurant advertising expenditure, and conduct a longitudinal analysis to determine how advertising has changed over time and its impact in counties with more low-income and/or Black and Hispanic residents. Stratifying by population density, and controlling for county-level effects, our results indicate that high dense counties, with lower income and higher minority populations, experienced larger increases in advertising expenditures. Using BMI data for 2.3 million adults from AthenaHealth and advertising exposure data from Nielsen Ad Intel, we also measure the impact these restaurant chains have on obesity risk via a population-based longitudinal model.

8c. Bayesian Joint Modeling of Multivariate Longitudinal and Survival Outcomes Using Gaussian Copulas

Seoyoon Cho, University of North Carolina at Chapel Hill

There is an increasing interest in the use of joint models for analysis of longitudinal and survival data. While random effects models have been widely used, these models can be hard to implement and can be computationally demanding. Copulas provide a useful alternative framework for joint modeling. One advantage of using copulas is that practitioners can directly specify marginal models for the outcomes of interest. We develop a joint model using a Gaussian copula to characterize the association between multivariate longitudinal and survival outcomes. Rather than using an unstructured correlation matrix in the copula model to characterize dependence structure as is common, we propose a novel decomposition that allows practitioners to impose structure (e.g., auto-regressive) which provides efficiency gains in small to moderate sample sizes and reduces computational complexity. We develop a Markov Chain Monte Carlo model fitting procedure for estimation. We illustrate the method's value using a simulation study and present a real data analysis of longitudinal quality of life and disease-free survival data from an International Breast Cancer Study Group trial.

8d. Model Selection and Inference in Variational Longitudinal Distributed Lag Models for Analyzing Post-flight Effects of In-flight Exposures

Mark Meyer, Georgetown University

Flight-related health effects are a growing area of environmental health research but one understudied area is on the post-flight effects of in-flight exposures. Studies investigating flight-related health effects often collect a range of repeatedly sampled, time-varying exposure-related measurements under both crossover and longitudinal sampling designs. A natural choice to model the relationship between these lagged exposures and post-flight outcomes is the distributed lag model (DLM). However, longitudinal DLMs (LDLM) are a lightly studied area. Thus, we propose an LDLM where the random effects can incorporate more general structures - including random lags - that arise from repeatedly sampling lagged exposures. We develop variational Bayesian algorithms to estimate LDLMs, derive a novel variational AIC, and show that the variational estimates can be used to test for the difference between two semiparametric curves under the crossover design. We then analyze the impact of in-flight, lagged exposure-related physiological effects on post-flight heart health. We also perform simulation studies to evaluate the operating characteristics of the LDLM and inference procedures.

8e. Considerations for Estimating Longitudinal Change with an Application to a Remote Unsupervised Digital Cognitive Battery

Roland Brown, Biogen

Data collected from digital health tools are increasingly common in clinical trials and observational studies and allow measurement of disease state in real-world settings at more frequent intervals than traditional measurement approaches. Common scientific questions in these studies involve quantifying longitudinal change in digital measures over the course of the study, and often comparing magnitude of change between groups. A variety of modeling choices are available to statisticians, including (1) modeling data on the original scale versus change from a baseline value; (2) selection of the baseline timepoint to properly account for learning effects; and (3) modeling temporal effects using categorical, linear, or higher-order terms. Each choice comes with differing assumptions and implications, and no clear consensus exists as to best practices. We illustrate the consequences of different modeling choices using simulated data and provide recommendations for practitioners. An example is provided using data collected from a digital cognitive battery in a remote, longitudinal cognitive health study.

9. POSTERS: MACHINE LEARNING AND COMPUTATIONAL METHODS

9a. P-Value Computation for Higher Criticism Tests

Wenjia Wang, University of Pittsburgh

In modern data analysis for detecting rare and weak signals, higher criticism (HC) and its variations have been an effective group-testing method with asymptotic optimality. Computation accuracy and speed, however, have long been an issue when the number of p-values combined or the number of tests are large. Several methods have been developed while they all have significant restrictions, especially when stringent significance level is required for multiple comparison correction. To this end, we propose two complementary approaches for HC and its four variations: (1) a cross-entropy based importance sampling method to provide accurate p-value calculation for any supremum domain of HC; (2) an analytic approach that alleviates numerical error. These two methods are integrated to form the final proposal for general HC p-value computation. Finally, an effective approach combining pre-calculated statistical tables and interpolation provides an ultra-fast computation O(1). Extensive simulations are implemented to benchmark accuracy and speed of proposed methods. Application to COVID-19 disease surveillance confirms its viability for large-scale studies.

9b. Neural-Network Transformation Models for Counting Processes

Rongzi Liu, University of Florida

The Cox model and the proportional odds model are the most popular ones in survival models. Both models are special cases of the linear transformation model. Non-linear functional form can also be specified in the linear transformation model. Nonetheless, the underlying functional form is unknown and mis-specifying it leads to biased estimates and reduced prediction accuracy of the model. To address this issue, we develop a neural-network transformation model. Similar to neural networks, the neural-network transformation model uses its hierarchical structure to learn complex features from simpler ones and is capable of approximating the underlying functional form of covariates. It also inherits advantages from the linear transformation model, making it applicable to both time-to-event analyses and recurrent event analyses. Simulations demonstrate that the neural-network transformation model outperforms the linear transformation model in terms of estimation and prediction accuracy when the covariate effects are non- linear. The advantage of the new model over the linear transformation model is also illustrated via two real applications.

9c. Novel Statistical Approaches to Address Dataset Shift in EHR-based Risk Prediction Models

Likhitha Kolla, Perelman School of Medicine, University of Pennsylvania

Reliable machine learning tools built on EHR data have the potential to forecast medical needs and advance clinical decision-making. However, it is widely recognized that these algorithms can deteriorate over time due to dataset shift. Dataset shift occurs when the distribution of the training data is different from that of the data from the deployment setting, resulting in reduced model performance. Common causes include changes in healthcare utilization and policies, which affect the data generation process. Given the high stakes setting of healthcare, proactively identifying shift, and adjusting models to maintain reliability is crucial. Methodology to detect and correct for shift is still nascent and largely relies on model retraining, which is computational expensive and leads to a loss of historical information. To this end, we evaluate novel statistical approaches to identify the drivers and extent of shift, as well as approaches to model maintenance, in a nationally implemented risk prediction model. Findings from this work will contribute to standardized evaluations of shift in EHR models, and discussions on the safe use of medical algorithms.

9d. Transportability of HIV Virologic Failure Prediction Models Across Nine Care Programs

Allan Kimaina, AMPATH

Routine viral load (VL) measurement is the gold standard for monitoring HIV treatment effectiveness. Efforts to identify patients likely to fail virologically have led to the development of several novel virologic failure (VF) risk predictive models. However, what remains understudied is the ability to transport such models from one health system to another. This study examines the cross-care-program transportability of machine learning models across nine care programs in 3 countries. Evaluation is done by training models using clinical data for each program and validating the accuracy of predictions when applied to other programs. To benchmark accuracy, we cross-validated (CV) a model that combined data across all programs. Data from 38,268 patients were included; 32,719 (85.5%) were virally suppressed at the first VL measurement. Average cross-care-program AUC (CI) across the nine care programs is 0.71 (0.69-0.72). CV AUC for the pooled data is 0.82 (0.80-0.84). In conclusion, we are finding that when an algorithm is derived from one population, discrepancies in the distribution of individual predictors substantially impact accuracy when applied to an external population.

9e. Bayesian Nonparametric Survival Analysis with Risk Set Adjustment for Left Truncation

Nikolay Krantsevich, Foundation Medicine, Inc.

Electronic health records (EHR) have led to increasing availability of survival data with the potential for rapid information accumulation. These data bring distinct challenges, including observational treatment assignment and large numbers of candidate effect modifiers. When EHR data are linked to other information sources like genomic test results, the event that triggers eligibility for the linkage can induce left truncation for time-to-event outcomes. While flexible machine learning approaches like Bayesian Additive Regression Trees (BART) can guard against residual confounding from misspecification, few of these accommodate any form of left truncation. We introduce an extension of time-to-event BART that accounts for independent delayed entry. We reduce execution times by providing the fully Bayesian algorithm with starting values derived from a faster tree construction during warmup. We demonstrate the utility of the method with a comparative effectiveness analysis in a linked clinicogenomic database of cancer patients.

9f. A Kernel Neural Network with a Computationally Efficient Variance Least Square Estimator

Heng Ge, University of Florida

Linear mixed model (LMM) and its extensions have long been the state-of-the-art choice for genetic risk prediction analysis of complex diseases. However, LMM commonly assumes a linear genotype-phenotype relationship, which may not be satisfied for diseases involving complex genetic etiology (e.g., interactions). Moreover, it remains computationally challenging to model a large number of samples, especially with multiple kernels built on genetic data. To address these rising challenges, we propose a kernel neural network (KNN) method with a variance least square estimator (VLS). The new method uses the hierarchical structure of kernel neural networks to model complex genotype-phenotype relationships (e.g., nonlinear and non-additive relationships), and uses VLS to improve computational efficiency. We further extend VLS to generalized VLS (GVLS), which can accommodate a variety of kernel neural network structures. Through simulations and a real data application, we have shown that KNN with VLS or GVLS outperformed current state-of-the-art methods in terms of accuracy and computation efficiency.

9g. FastQDesign: A Realistic FASTQ-Based Framework for ScRNA-Seq Study Design Issues

Yu Wang, Medical College of Wisconsin

Single-cell RNA sequencing has emerged as a powerful tool for characterizing transcriptomic profiles at single-cell resolution. Designing such experiments not only requires considering the number of cells but also sequencing depth, which is a two-dimensional optimization problem when given a fixed budget. Existing literature addresses this problem using simulation-based approaches and only considers the UMI matrix as input. Although the UMI matrix is the standard input of many analysis pipelines, it is also known to be affected by technical factors such as sequencing saturation. Here we proposed a design framework that utilizes raw fastq files of reference dataset for the design, namely 'FastQDesign.' Our approach is based on the rarifying procedure for the reference dataset and searching for the optimal design that achieves the highest stability against rarification by evaluating cluster co-membership and cluster marker genes. We demonstrated our framework using a NOD mice T-cell dataset and performed the cost-benefit analysis by identifying the optimal design.

9h. On Estimation of Radiation Dose Response Parameters

Brian Egleston, Fox Chase Cancer Center

Estimating parameters of radiation dose response models is of interest. We present the Lyman-Kutcher-Burman model and discuss how it can be estimated. We describe a simulation examining a four parameter model that incorporated the alpha/beta ratio, and also a six parameter model that added time latency effects. We demonstrate convergence and operating characteristics when we start estimation at the truth and at the default settings. Next, we present a didactic data example of a phase 3 clinical trial in which we estimate parameters using a grid search.

We find that estimators of the Lyman-Kutcher-Burman model have convergence difficulties. The finite sample distributions of some estimators are skewed. This can result in unstable and highly variable estimates of the parameters. In addition, the model is prone to convergence on local rather than global maximum likelihood estimates. The highly non-linear model can cause substantial variability in estimates across studies. Authors who estimate the model could provide more details on the convergence properties of their estimates, and ensure that estimates reflect global rather than local maximum likelihood values.

9i. Machine Learning Approach for Predicting Cognitive Impairment and Identifying Risk Factors in Elderly People

Monsuru Durojaiye, Austin Peay State University

Alzheimers Disease (AD) is one of the most diagnosed diseases among elderly patients. AD is a progressive neurodegenerative disease characterized by memory disfunction and cognitive decline. Cognitive decline, a natural process of aging may develop into cognitive impairment. It has been observed that elderly patients with serious cognitive impairment have a higher risk of progressing into AD. Due to lack of effective treatment for AD, it is important to emphasize on prevention and early identification signs of patients with risk of cognitive impairment. In our research, we focus on the elderly people with or at risk of cognitive impairment. Several studies have predicted that different factors such as brain imaging, biomarkers, demographic etc. show potential progression of cognitive impairment to AD. However, models that use brain imaging and biomarkers are very costly. Also, most of the factors are not modifiable. Therefore, to fill these gaps, in this study we aim to build prediction models using machine learning algorithms to early identify the elderly at risk for cognitive impairment in advance.

9j. Single-Cell Linear Adaptive Negative-Binomial Expression Testing (scLANE)

John Leary, University of Florida

Single cell RNA-sequencing (scRNA-seq) offers a high-resolution view of cellular biology, including dynamic processes such as differentiation and disease progression. Many methods have emerged that estimate a cell-level time ordering from static scRNA-seq samples, which use similarity of gene expression to place cells in order on some biological manifold. Researchers then typically 1) assume the ordering represents a biological process and 2) characterize which genes are associated with that process. Changes in expression over trajectories are usually complex with generalized additive models (GAMs) the dominant current choice of model. However, while GAMs are excellent for fitting nonlinear relationships, they are not easily interpretable. To address this trade-off, we developed single-cell Linear Adaptive Negative-binomial Expression (scLANE) testing. scLANE balances the need for a flexible nonlinear model and the facilitation of biological interpretation. We demonstrate our method’s accuracy and ability to draw meaningful comparisons from the model on simulated data and a case-study datasets having tens of thousands of cells and from multiple subjects.

9k. Asymptotic Properties of Deep Neural Network Sieve Estimators

Chang Jiang, University of Florida

Deep Neural Networks (DNN) attain great success in various disciplines (e.g., imaging recognition, natural language processing, and genomics research). While many studies have focused on bounding the prediction error of neural network estimators to explain the remarkable prediction performance of DNN, limited research has been conducted on the statistical inference of DNN. In this article, we study the sieve estimator of feed-forward DNN in a nonparametric framework. We show that the sieve estimator is consistent and able to achieve a good rate of convergence even in the high-dimensional setting. Moreover, a novel hypothesis test is proposed on a known function with smoothness conditions in the DNN model. The test statistic is easy to implement and follows an asymptotic normal distribution. The theoretical result is also verified via simulation studies and real data applications.

9l. Identification of Key Chemical Sub-Structures in Protein-Compound Binding Using a Sparse Graph Neural Network Model

Zanyu Shi, Indiana University Fairbanks School of Public Health

Many protein kinases have been recently identified as promising inhibitor targets for Alzheimer's disease drug therapy. Deep learning methods have been successfully applied to predict protein-compound affinity and virtual screening for drug discovery. However, they majorly focus on prediction accuracy and do not provide insights into drug design. To fill the gap, we propose a sparse graph neural network (GNN) model that not only can predict protein-compound affinity but also can identify essential molecular substructures in protein-compound binding. We apply generalized group Lasso to enforce the GNN model to prune non-essential substructures in the compounds but promote the ones that are key to protein-compound binding. We compare our sparse GNN model with the state-of-the-art interpretable deep learning model on the binding perdition between drugs and tyrosine-protein kinase LYN. We find that our spare GNN model outperforms the competing models in terms of perdition accuracy. Furthermore, the chemical substructures identified by our model tend to have lower binding energy, suggesting their importance for protein-compound binding.

9m. Neyman-Pearson and Equal Opportunity: When Efficiency Meets Fairness in Classification

Shunan Yao, University of Southern California

Organizations often rely on statistical algorithms to make socially and economically impactful decisions. We must address the fairness issues in these important automated decisions. On the other hand, economic efficiency remains instrumental in organizations' survival and success. Therefore, a proper dual focus on fairness and efficiency is essential in promoting fairness in real-world data science solutions. Among the first efforts towards this dual focus, we incorporate the equal opportunity (EO) constraint into the Neyman-Pearson (NP) classification paradigm. Under this new NP-EO framework, we (a) derive the oracle classifier, (b) propose finite-sample based classifiersthat satisfy population-level fairness and efficiency constraints with high probability, and (c) demonstrate statistical and social effectiveness of our algorithms on simulated and real datasets.

9n. WRdesign: A Web App for Exploring Win Ratio Designs

Li Huihua, Pfizer Inc

The win ratio is an innovative statistical method for analyzing composite endpoint. Despite its rising popularity, sample size calculation exists as a major obstacle in designing trials with the win ratio due to intricacy from the composite construction, correlations between components and computational complexity in simulation. We present Win Ratio Design (WRdesign) web app for accelerating sample size calculation and allowing a fast and comprehensive evaluation of a win ratio study at the design stage. WRdesign employs a novel efficient win ratio algorithm, fast simulation-based binary search for sample size determination, and comparisons with alternative formula-based approaches. A robust back-end data structure provides flexibility in specifying the composite endpoint from a broad range of different outcome types and correlation among components. WRdesign makes evaluating the sensitivity of power to changes in assumptions and understanding the relative efficiency of added components straightforward. WRdesign can be used to explore design aspects required for optimizing win ratio performance, leading to selection of a compelling and clinically meaningful trial outcome.

9o. Simulation Models for Bladder Cancer: A Systematic Review

Timothy Hedspeth, Brown University

We searched electronic databases for English language papers, including combinations of terms related to bladder cancer; and (micro)simulation, computer-based model, state or dynamic transition, risk prediction, discrete event simulation, cohort-/population- based model. We found models employing differential equations for the BC biology, Decision Trees for CEAs, while more complicated cohort-based models, incorporating bidirectional transitions between Markov states and Microsimulation models are used to describe more complex systems and assess screening and treatment protocols on BC incidence and mortality, accounting for recurrence. Key outcomes include diagnosis, recurrence, and mortality with most commonly BC types the NMIBC and MIBC. Commonly evaluated treatments are BCG immunotherapy, TURBT, Mytomycin C, chemotherapies, and cystectomy, or combinations of those. The wide use of simulation models makes evident the potential these approaches have to explore important facets of the disease. This systematic review revealed the need for good reporting practices, so as researchers can benefit from simulation studies to improve and enhance decision making for bladder cancer.

9p. Efficient Computation of High-Dimensional Penalized Generalized Linear Mixed Models by Latent Factor Modeling of the Random Effects

Hillary Heiling, University of North Carolina at Chapel Hill

Modern biomedical datasets are increasingly high dimensional and exhibit complex correlation structures. Generalized Linear Mixed Models (GLMMs) have long been employed to account for such dependencies. However, proper specification of the fixed and random effects in GLMMs is increasingly difficult in high dimensions, and computational complexity grows with increasing dimension of the random effects. We present a novel reformulation of the GLMM using a factor model decomposition of the random effects, enabling scalable computation of GLMMs in high dimensions by reducing the latent space from a large number of random effects to a smaller set of common factors. We also extend our prior work to estimate model parameters using a modified version of the Monte Carlo Expectation Conditional Minimization algorithm, allowing us to perform variable selection on both the fixed and random effects simultaneously. We show through simulation that through this factor model decomposition, our method can fit high dimensional penalized GLMMs faster than comparable methods and more easily scale to larger dimensions not previously seen in existing approaches.

9q. Construction of Latin Hypercube Design and its Evaluation

Arata Ueda, Keio University

Latin Hypercube Designs (LHD) is one of the space filling designs and widely used in many filed. As the number of experiments and factors increases, the total number of LHDs increases. Because of this, it is difficult to construct optimal LHD. To overcome this problem various algorithms for constructing optimal LHDs have been proposed in the previous studies. In this study, we propose a algorithm to construct LHDs based on the Artificial Bee Colony algorithm and a combined algorithm, which is combining the Artificial Bee Colony algorithm and Differential Evolution algorithm. In addition, we propose an algorithm that adds a process to get out of the local optimum. we yield the good LHDs from the proposed algorithms compared to the LHDs yielded from the algorithms proposed in the previous studies in some cases. Moreover, we compare LHDs not only by the conventional design criterion, such as extended maximin distance criterion and ?q criterion, but also by the cumulative frequency of the distances between the design points. This comparison gives a new aspect of the goodness of LHD.

10. POSTERS: MISSING DATA AND MEASUREMENT ERROR

10a. Estimators for Longitudinal Mixed Effects Models to Account for Right Censored Predictors

Jesus Vazquez, University of North Carolina at Chapel Hill

Success of clinical trials aimed at slowing down the progression of Huntington&rsquo;s Disease (HD) depends on correctly modeling of how symptoms progress, especially before a clinical diagnosis where therapies have the best chance of slowing symptoms. This is problematic because time to clinical diagnosis is not always observed: patients drop out or the study ends which leads to a censored value for time to clinical diagnosis. To address this problem, the analysis is often restricted to observations with only fully observed data but this is often inefficient since estimates will have higher variability due to the reduction in sample size. We illustrate how weighting estimation, imputation, two-stage likelihood methods compare and under what assumptions each of these methods will correctly model the progression of HD when age at clinical diagnosis is censored. We apply our method to a study of HD to model the progression of symptoms.

10b. Sensitivity Analysis for Non-ignorable Missing Data in Sequential Multiple Assignment Randomized Trials: A Delta-Adjusted Controlled Imputation Approach

Aparajita Sur, University of Minnesota

Sequential multiple assignment randomized trials (SMARTs) are increasingly needed to develop adaptive interventions for disorders with heterogeneous treatment effects. However, missing data can compromise the validity of inference and the complex structure of SMARTs presents unique challenges when handling missing data. While multiple imputation can facilitate unbiased and efficient estimation for longitudinal data, it relies on the unverifiable assumption that the data is missing at random (MAR). It is unclear how violations of the MAR assumption affect inference in a SMART setting, highlighting the need for sensitivity analyses to assess the robustness of conclusions to departures from MAR. However, existing sensitivity analysis methods for longitudinal data do not address the structural missingness unique to a SMART design. We propose a flexible delta-adjusted controlled imputation framework to perform sensitivity analyses in SMARTs amidst both monotone and non-monotone missingness. We validate our approach with a simulation study and implement our framework to explore how departures from MAR affect conclusions in a SMART study addressing college binge drinking.

10c. Jackknife Variance Estimator for Datasets Containing Multiply Imputed Outcome Variables Under Uncongeniality: A Monte Carlo Simulation Study

Ihsan Buker, University of West Florida

Missing data is an issue ubiquitous in statistics. Today, multiple imputation (MI) is one of the most commonly utilized approaches to provide valid statistical inferences in the presence of missing data. Accompanying MI is the issue of uncongeniality, which occurs when the imputation model and the analysis model make different assumptions. A set of rules proposed by Rubin to pool parameter estimates was shown to produce biased point estimates under uncongeniality and either conservative or anti-conservative variance estimates.

Combined MI and resampling methods have been proposed as robust estimators. Bootstrapping, one of the most commonly utilized resampling methods alongside MI to obtain proper estimates, has its basis in asymptotic theory. As such, the need for a robust estimator remains in small samples frequently encountered in biological studies. We propose a jackknife estimator for small, multiply imputed datasets, the performance of which is investigated using a Monte Carlo simulation study. Accordingly, the recommendation is made to replace Rubin's rules as the de facto standard. An implementation of the proposed jackknife variance estimator in R is provided.

10d. Improving Regression Analysis with Imputation in a Longitudinal Study of Alzheimer's Disease

Ganesh Chandrasekaran, University of Pennsylvania

Missing data is prevalent in the Alzheimer's Disease Neuroimaging Initiative (ADNI) data set. Biological samples from cerebrospinal fluid, for instance, are missing measurements from half the subject database. It is common to deal with such missing data by removing subjects with missing entries prior to statistical analysis; however, this can lead to significant efficiency loss and possibly even bias. While previous studies have considered the problem of missing data in ADNI, it has yet to be demonstrated that the imputation approach to handling this issue can improve statistical efficiency in the longitudinal regression setting. Accordingly, the purpose of this study is to demonstrate the importance of proper imputation in ADNI by analyzing longitudinal cognitive exam scores and their association with baseline patient characteristics. We show that application of the MICE algorithm leads to valid, tighter confidence intervals, thus improving the analysis efficiency. Application of MICE is justified by investigating the missing data mechanism and model assumptions. We then assess robustness of results to the choice of imputation method and post-processing steps.

11. POSTERS: MULTIVARIATE AND SURVIVAL METHODS

11a. Genome Hierarchy Grouping Structure and Correlation Guided Feature Selection for Multivariate Outcome Prediction

Xueping Zhou, University of Pittsburgh

Developing efficient feature selection and accurate prediction algorithms for multivariate phenotypes is a major and difficult task in analyzing omics data. Many methods have been proposed to perform such tasks for univariate outcome, including top-performing penalized regression-based approaches. In this study, we propose a supervised learning algorithm to perform feature selection and multivariate outcome prediction for data with potentially high-dimensional predictors and responses. The method incorporates known genome hierarchy grouping and correlation structures into feature selection, regression coefficient estimation, and outcome prediction under a penalized multivariate multiple linear regression model. Extensive simulations show its superior performance. We apply the proposed method to a multi-omics dataset with two applications. In the first study, it achieves better cell type fraction prediction using bulk RNA-seq data. In the second association study between multivariate gene expression and high-dimensional DNA methylation data, it reveals novel association signals, providing insights on how CpG sites regulate gene expressions.

11b. Probabilistic Index for Survival Analysis Using Landmark Times, Clinically Significant Differences, and Multiple Prioritized Endpoints: A Simulation Study

Kanako Fuyama, Hokkaido University

Pairwise comparison methods are increasingly used to analyze survival outcomes in randomized controlled trials. These estimation methods and their corresponding probabilistic indices have been given various names by different authors: generalized pairwise comparisons, the net chance of longer survival, the restricted chance of longer survival, the net benefit, and the win ratio. The main idea of these frameworks is to consider the probability that a subject in the experimental arm lives longer than a subject in the control arm. Recent advances in such frameworks include the incorporation of landmark times and clinically significant differences and the comprehensive evaluation of multiple prioritized endpoints. In this simulation study, we visualize how the landmark times, clinically significant differences, and correlations between the prioritized outcomes impact the probabilistic indices using realistic clinical scenarios.

11c. Marginal Proportional Hazards Models for Multivariate Interval-Censored Data

Yangjianchen Xu, University of North Carolina at Chapel Hill

Multivariate interval-censored data arise when there are multiple types of events or clusters of study subjects such that the event times are potentially correlated and when each event is only known to occur over a particular time interval. We formulate the effects of potentially time-varying covariates on the multivariate event times through marginal proportional hazards models while leaving the dependence structures of the related event times unspecified. We construct nonparametric pseudo-likelihood under the working assumption that all event times are independent and present a simple and stable EM-type algorithm. The resulting nonparametric maximum pseudo-likelihood estimators for the regression parameters are shown to be consistent and asymptotically normal, with a limiting covariance matrix that can be consistently estimated by a sandwich estimator under arbitrary dependence structures for the related event times. We evaluate the performance of the proposed methods through extensive simulation studies and provide an application to data from the Atherosclerosis Risk in Communities Study.

11d. Variance Component Score Test for Multivariate Change Point Detection

Melissa Martin, University of Pennsylvania

Multivariate change point detection is a useful tool for detecting distributional changes in time-ordered data, but challenges arise when there is high dimensionality in the number of features relative to the number of observations. This problem is often encountered in mobile health, where we wish to detect significant changes in behavior in a high dimensional feature vector constructed from multi-modal data in patients at-risk for adverse events (e.g. relapse), as we may use behavioral changes to prompt timely interventions. Thus, powerful multivariate change point detection methods are needed to identify significant behavioral changes prior to the occurrence of these adverse events. Existing methods suffer from biased estimates of mean and covariance matrices when change points occur. To avoid this, we propose a variance component score test (VC*) that leverages only pre-change point data in a regularization framework to estimate the feature mean vector and covariance matrix. We compare VC* with existing methods in various simulation settings; we also apply these methods to smartphone sensor data collected in patients with schizophrenia to predict relapse and hospitalization.

11e. A Multivariate Beta Mixture Model Approach to Examining Racial/Ethnic and Socioeconomic Disparities in Endometrial Cancer Care in Massachusetts

Carmen Rodriguez Cabrera, Harvard T. H. Chan School of Public Health

Endometrial cancer(EC) is the most common gynecologic cancer in the United States, affecting 1 in 37 women each year. On average, African American women have 55% higher 5-year mortality risk compared to white women,and like other minority groups,they are vulnerable to receiving suboptimal care, likely due to difficulties accessing care stemming from the socioeconomic environments in which they reside.Previous research has examined socioeconomic factors(e.g., education, income) individually/independently,but these often interact as social determinants of health.Through implementation of a multivariate beta mixture model(MBMM), we aim to identify clusters of social determinant profiles using several racial-ethnic and socioeconomic factors.Using census tract aggregate level data and patient-level information from 11287 patients collected in the 2006-2017 Massachusetts Cancer Registry,we will identify differences in receipt of optimal care for EC patients in Massachusetts by social determinant cluster profiles. We will also compare the stability of our cluster profiles across three waves of the American Community Survey 5-year estimates(2006-2010,2011-2015,and 2015-2019).

11f. Variable Selection in Presence of Interaction Effects of Correlated High Dimensional Covariates Under Strong Heredity Constraints with Discrete Survival Frailty Model with Application to Mixtures of Environmental Toxicants and Time-to-Pregnancy

Abhisek Saha, NICHD, NIH

As modern environmental studies increasingly focus on the impact of toxicants on reproductive health, understanding the association between toxicants and time-to-pregnancy (TTP) becomes an important scientific goal. Assessing mixtures of chemicals' effects on TTP poses significant statistical challenges: 1. TTP being a discrete survival outcome subject to left truncation and right censoring, 2. highly correlated exposures 3. accounting for chemicals binding to lipids, and 4. High proportion of chemical values below the limit of detection. Although the authors address them in earlier works, incorporating interaction effects remained a challenge. We bridge the gap by proposing novel discrete survival frailty modeling along with an efficient algorithm that allows selection of interaction effects subject to strong heredity constraints while addressing the above issues. We conducted simulations under a variety of scenarios including the truth being hierarchical, anti-hierarchical and non-interaction, to show that our method performs better in terms of FP and FN rates compared to L1-based alternatives. We are currently illustrating the method in the LIFE Study.

11g. Fitting the Cox Proportional Hazards Model to Big Data

Jianqiao Wang, University of North Carolina at Chapel Hill

We propose a computationally efficient method for fitting the Cox proportional hazards model to big data involving millions of study subjects. Specifically, we perform maximum partial likelihood estimation on a small subset of the whole data and improve the initial estimator by incorporating the remaining data through one-step estimation with estimated efficient score functions. We show that the final estimator has the same asymptotic distribution as the conventional maximum partial likelihood estimator using the whole dataset but requires only a small fraction of computation time. We demonstrate the usefulness of the proposed method through extensive simulation studies and analysis of the UK Biobank data.

11h. Pseudo-Value Regression of Clustered Current Status Data with Informative Cluster or Subcluster Sizes in a Multistate Model

Samuel Anyaso-Samuel, University of Florida

Current status data presents a more severe form of censoring due to the single observation of study units transitioning through a sequence of well-defined disease states at random inspection times. The current status data may be clustered within specified groups, and informativeness of the cluster sizes may arise due to a relationship between the transition outcomes and the cluster sizes. Failure to adjust for this informativeness will lead to a biased inference. Motivated by estimating covariate effects on the state occupation probability (SOP), we propose extending the pseudo-value approach to clustered current-status data with informative cluster or subcluster sizes. First, we formulate the pseudo-values by marginal estimators of the SOP computed using nonparametric regression theory. Secondly, the estimating equations based on the pseudo-values are reweighted by functions of the cluster sizes to adjust for informativeness. We perform simulation studies to investigate the pseudo-value regression under different scenarios of informativeness. The method is applied to a motivating periodontal disease dataset which encapsulates the complex data-generating mechanism.

12. POSTERS: OMICS AND BIOMARKERS

12a. A Flexible Quasi-Likelihood Model for Microbiome Abundance Count Data

Yiming Shi, Washington University School of Medicine in St. Louis

In this paper, we present a flexible model for microbiome count data. We consider a quasi-likelihood framework, in which we do not make any assumptions on the distribution of the microbiome count except that its variance is an unknown but smooth function of the mean. By comparing our model to the negative binomial generalized linear model (GLM) and Poisson GLM in simulation studies, we show that our flexible quasi-likelihood method yields valid inferential results. Using a real microbiome study, we demonstrate the utility of our method by examining the relationship between adenomas and microbiota.

12b. Statistical Considerations in Differential RNA Methylation Data Analysis

Daoyu Duan, Case Western Reserve University

Various types of RNA methylation, including N6-methyladenosine (m6A), are involved in human disease development. MeRIP-seq, a newly developed sequencing biotechnology, can quantify the m6A level on a transcriptome-wide scale. One of the fundamental questions in RNA methylation data analysis is to identify the Differentially Methylated Regions (DMRs). Multiple statistical approaches have been developed for DMR detection over the last couple of years. Here, we investigate and benchmark them, using both synthetic and real data. First, we thoroughly assess all eight existing methods for DMR calling. Next, we design a hierarchical model in-silico simulator to generate new m6A data. Last, we provide a statistical power assessment tool for researchers who want to conduct differential RNA methylation analysis in research.

12c. Integrative Metagenomics and Metatranscriptomics Statistical Analysis Using GMCM in Human Microbiome Data

Chuwen Liu, University of North Carolina at Chapel Hill

Human microbiome are key factors in many aspects of health problems. Metagenome and Metatranscriptome, usually measured by DNAseq and RNAseq, separately capture the abundance of microbial species and the functional characterization of microbes. However, integration of the paired DNAseq and RNAseq data is still under development. One goal to be achieved by such integration is to identify the differential relative expression (DrE, as logarithm of the ratios between normalized counts in RNAseq and DNAseq) activities at the species or gene level. The distribution of Log Ratios may have >1 mode due to excess zero typically in microbiome data. Therefore, a Gaussian Mixture multivariate Regression with Concomitant variable Model (GMCM) is proposed to fit the Log Ratios after removing samples with zero DNAseq counts. Our proposed method has the advantage of handling multiple component-specific microbiome hypotheses simultaneously with batch effects controlled. Simulations show the controlled false positive rate and good power. The application in studying Early Childhood Caries and Inflammatory bowel diseases identified disease-associated DrE species/genes.

12d. Improving Precision and Power in Detecting Differentially Methylated Regions

Daniel Alhassan, Missouri University of Science and Technology

Detecting differentially methylated regions (DMRs) between different biological conditions is critical for identifying disease biomarkers. Though methods for detecting DMRs in microarray data have been introduced, developing methods with high precision, power, and accuracy in determining the true length of DMRs remains a challenge.

One popular method accounts for the correlation between nearby CpG sites by using probe spacing on the array. In this study, we used a normalized kernel-weighted model to account for co-methylation using the relative probe distance from nearby CpG sites. When compared to a popular DMR detection method, our method consistently has higher power and precision in detecting a true effect. Furthermore, our method is more accurate in determining the true DMR length. In terms of the number of significant DMRs detected, our findings were consistent when applied to oral cancer data. We also found that methylation differences occurred in the Akt1 and 2 genes, which are usually overexpressed in oral cancer tissues. With the substantial increase in precision and power, reliable disease biomarkers can be identified.

12e. Accurate Estimation of Rare Cell Type Fractions from Tissue Omics Data via Hierarchical Deconvolution

Penghui Huang, University of Pittsburgh

Gene expression in tissue samples reflects the average expression levels across different cell types and is highly influenced by cellular fractions. As such, cellular fractions may confound tissue-level analyses. Dozens of cellular deconvolution methods have been developed to resolve this issue. However, existing methods are designed for tissues consisting of clearly distinguishable cell types and have di?iculties estimating highly correlated cell subtypes. To address this challenge, we propose Hierarchical Deconvolution (HiDecon), a penalized approach that uses single-cell RNA sequencing data to estimate cellular fractions in tissue samples guided by a hierarchical tree modeling the similarities among cell types and cell differentiation relationship. By coordinating cell fractions across layers of the hierarchical tree, cellular fraction information is passed upwards and downwards along the tree, which helps correct the estimation bias. We can also estimate rare cell fractions via splitting the tree to a high resolution. Through simulations and real data with ground truth of measured fractions, we demonstrate that HiDecon significantly outperforms existing methods.

12f. DiffCircaPipeline: A Framework For Multifaceted Differential Rhythmicity Analysis

Xiangning Xue, University of Pittsburgh

Circadian oscillations of gene expression regulate daily physiological processes, and their disruption is linked to many diseases. Circadian rhythms can be disrupted in a variety of ways, including differential phase, amplitude, and rhythm fitness. Although many differential circadian biomarker detection methods have been proposed, a workflow for systematic detection of multifaceted differential circadian characteristics with accurate false positive control is not currently available. We propose a comprehensive and interactive pipeline to capture the multifaceted characteristics of differentially rhythmic biomarkers. Analysis outputs are accompanied by informative visualization and interactive exploration. The workflow is demonstrated in a real world study and is extensible to general omics applications.

12g. Data-Driven Evaluation of Trajectories in Single-Cell RNA-Seq Data

Xiaoru Dong, University of Florida

Trajectory inference (TI) is a widely applied analysis to single-cell RNA-sequencing data in which single-cells are computationally ordered or graphed based on their expression profiles. Consequently, TI is often employed to model dynamic cellular changes even in single time-points datasets. While many tools currently exist to order the cells, a number of preliminary tasks such as feature selection and dimension reduction are important for pre-processing. We demonstrate the profound affect these choices have on the reliability of the trajectory estimation. To address this challenge in analyzing scRNA-seq data, we developed a two-stage evaluation framework to firstly, detect the existence of trajectories and secondly, evaluate a trajectory's goodness-of-fit. We demonstrate the ability of our method to improve decision making in trajectory inference analysis on simulated and case-studies data.

12h. Evaluation of Differential Epitranscriptome Analysis Methods

Wen Tang, Case Western Reserve University

Various types of RNA methylation, including N6-methyladenosine (m6A), are involved in human disease development. As a newly developed sequencing biotechnology to quantify the m6A level on a transcriptome-wide scale, MeRIP-seq expands RNA epigenetics study in both basic and clinical applications. One of the fundamental questions in RNA methylation data analysis is to identify the Differentially Methylated Regions (DMRs). Multiple statistical approaches have been developed for DMR detection, but a comprehensive evaluation of these methods is lacking. Here, we thoroughly assess all eight existing methods for DMR calling, using both synthetic and real data. Our simulation adopts a Gamma-Poisson model and accommodates various sample sizes and DMR proportions. TRESS and exomePeak2 perform the best using metrics of detection precision, FDR, type I error control, and runtime, though hampered by low sensitivities. DRME and exomePeak obtain high sensitivities, at the expense of inflated FDR and type I error. Analyses of three real datasets suggest differential preference on identified DMR length and uniquely discovered regions, between these methods.

12i. A Deep-Transfer Learning Based Model to Identify High-Risk Components of TNBC

Tianhan Dong, Indiana University

Triple negative breast cancer (TNBC) is the most severe breast cancer (BRCA) subtype. TNBC patients easily develop resistance to chemotherapies and lack therapeutic targets, it is urgent to develop effective precision medicine for TNBC patients. We advocate for the application of deep transfer learning techniques to TNBC samples to better understand what components correlate with these poor outcomes of TNBC. Using our newly developed Diagnostic Evidence Gauge of Single-cells (DEGAS) framework, we can identify high-risk cells and tissue regions in BRCA. We have also applied DEGAS to BRCA spatial transcriptomics (ST) datasets where we found tumor regions that associated with higher breast cancer risk. Cellular topology from corresponding H&E images was used to explore the underlying mechanisms among high-risk tumor regions. To test our models, four TNBC clinical samples were collected, processed, and submitted for ST and single cell RNA sequencing, to examine the power of our model on identifying high-risk components of TNBC. In conclusion, we are developing and applying deep-transfer-learning-based risk inference models for TNBC

12j. Identify Alzheimer's Disease Subtypes with Integration of Multiple Omics Data Using Subspace Merging Algorithm

Ziyan Song, Indiana University

Alzheimer's disease is a progressive brain disorder. Identifying subtypes of AD can successfully improve diagnosis, treatment, and future disease management. The goal of this project is to identify subtypes of AD by integrating multiple types of omics data. We applied a subspace merging algorithm on AD patients with matched gene expression, DNA methylation, proteomics data collected from the religious orders study and memory and aging project. Patient-to-patient similarity matrix was constructed for each data type, and then the corresponding subspace graph representations were calculated and merged on a Grassmann manifold. A spectral clustering method was applied to generate the clusters. Clinical and pathological scores were compared statistically among the clusters. Our results identified two patient clusters in which the cognitive test scores or the degree of pathology in AD are significantly different. Differential analyses were conducted on all three types of omics data and enrichment analysis was conducted to identify the underlying biological processes and pathways through which a group of genes highly significantly related to synaptic functions are identified.

12k. Comprehensive Cell-Type Deconvolution of Liquid Biopsy Samples Using Methylated Cell-Free DNA

Arthur McDeed, Georgetown University

Decoding the origin of cell-free DNA (cfDNA) can reveal altered cellular contributions reflective of changes in disease state, tissue specific damage, and treatment related adverse events. In this work, we develop a novel statistical method for performing methylation-based cell-type deconvolution utilizing an Expectation-Maximization algorithm. In contrast to prior methods using a single CpG site, our flexible, probabilistic method leverages the co-regulation of neighboring CpG sites, analyzing the methylation patterns of cfDNA fragments at the read level to facilitate multi-class deconvolution of liquid biopsy samples. In simulation studies, we demonstrate robustness of our model estimates under the range of read level methylation patterns and noise levels. We then assess the accuracy of cell-type proportion estimation on in-silico mixed cfDNA samples from real WGBS data. Finally, we apply our model to two real-world examples in which we show the ability of our model to detect off-target tissue damage in breast cancer patients undergoing radiation therapy and increased cellular contributions from leukocytes in ALL patients compared to healthy controls.

WITHDRAWN 12l. Decomposing Interaction and Mediating Effects of Race/Ethnicity and Circulating Cystatin C on Cognitive Status in the United States Health and Retirement Study

Cesar Higgins Tejera, University of Michigan, School of Public Health

Background: Elevated circulating cystatin C is associated with cognitive impairment in non-Hispanic Whites, but its role in racial disparities in dementia is understudied. Methods: In a cross-sectional sample of the Health and Retirement Study (n=9,921), we estimated prevalence ratios and to test the relationship between elevated cystatin C (>1.24mg/L vs <1.24mg/L) and impaired cognition. We calculated additive interaction measures and conducted four-way mediation-interaction decomposition analysis to test the moderating effect of race/ethnicity and mediating effect of cystatin C on the racial disparity. Results: Elevated cystatin C was associated with dementia (prevalence ratio [PR] = 1.4; 95%CI: 1.2, 1.8). Among non-Hispanic Black relative to non-Hispanic White participants elevated cystatin C was estimated to account for 2% (95% CI: -0%, 4%) for the racial disparity in prevalent dementia, and the interaction accounted for 9% (95% CI: -4%, 23%). Discussion: Cystatin C is associated with adverse brain health; this effect is larger than expected for individuals racialized as minorities had they been racialized and as non-Hispanic White.

12m. Joint Microbiome and Metabolome Association Analysis with a Non-Parametric Conditional Approach

Chang Chen, University of Washington

Studying the relationship between the microbiome and metabolome promises to advance our understanding of their interactions and aid in deciphering the complex mechanisms underlying microbial origins of complex diseases. A critical step in analyzing microbiome-metabolome relationships is the assessment of the association between individual bacterial taxa and metabolites. This typically proceeds via classical approaches such as simple pairwise Pearson or Spearman correlation. However, such measures are hindered by the high correlation among metabolites and among taxa, leading to excessive spurious associations. Therefore, we propose to leverage the Scaled Expected Conditional Covariance (SEcov), a nonparametric conditional association analysis, to recover direct associations between microbes and metabolites. Its key is the adjustment for all other taxa or metabolites when assessing the association between a taxon-metabolite pair which thus excludes indirect interactions. We conduct simulations to evaluate the performance of SEcov and compare it with other correlation-based techniques, demonstrating its high association recovery accuracy. We also illustrate SEcov in real data.

13. POSTERS: ORDINAL AND CATEGORICAL DATA ANALYSIS

13a. Expected Number of Stages for Hierarchical Group Testing Algorithms

Minh Nguyen, University of Nebraska - Lincoln

Group testing is widely used by laboratories for infectious disease detection. Its application involves testing pooled sets of clinical specimens, rather than initially testing individual specimens, to determine positive/negative outcomes for each person. When applied with judiciously chosen group sizes, group testing greatly decreases the number of tests needed for testing high volumes of clinical specimens. Prior to its implementation, the expected number of tests for a group testing algorithm needs to be calculated to assess an algorithm's efficiency. However, this metric does not account for the complexity of algorithm implementation. This is especially important to understand for hierarchical-based group testing algorithms that may take a large number of stages to complete. For this reason, we propose adopting the expected number of stages for an algorithm as a new measure of efficiency that can be used along with the expected number of tests. In our presentation, we examine its interpretation and outline its derivation for two and three-stage hierarchical algorithms. We illustrate our new metric in the context of recent applications of group testing for SARS-CoV-2 detection.

13b. Addressing the Impact of Group Imbalance on Standardized Effect Size Measures in Real-World Evidence Studies

Adam Sima, CorEvitas, LLC

Standardized effect sizes (ES) have become popular assessing the magnitude of association between two variables without relying on hypothesis testing. Omnibus standardized ES, such as Cramer's V, are available to measure the magnitude of association between two variables if one of these variables is multinomial or ordinal. These quantities are based on the Pearson chi-square statistic which is sensitive to the sample size balance across the different levels of the multinomial or ordinal variable. This work demonstrates how omnibus standardized ES based on imbalanced sample sizes are smaller in magnitude than similar estimates originating from balanced situations. We propose reporting omnibus standardized ES that preserve the treatment group association but equate the marginal distribution across groups. Creating balance across groups, achieved by weighting the sample or using expected values assuming balanced groups, is shown to be robust against imbalanced marginal distributions. Our motivation is drawn from the CorEvitas Psoriasis Registry, where we demonstrate our proposed methodology has a significant impact on the magnitude of the omnibus standardized ES.

13c. Estimating Principal Causal Effect for Multiple Levels of Noncompliance Using Principal Ignorability

Elizabeth Sarker, Johns Hopkins University

In this paper, we show the identification and estimation of the principal causal effects (PCE) for multilevel compliance using the Principal Ignorability assumption and apply it to the Baltimore Experience Corps trial. We also show the estimation of differential impacts across the compliance levels over common regions of covariates. Though it is possible to estimate the PCE for multilevel compliance, estimation becomes challenging as the number of levels increases in a finite sample and can lead to large variances. Also, such a stratification strategy does not allow for the incorporation of the scientific knowledge researchers may have about the compliance behavior of the study. In an experiment, it might be valid to assume that compliance and outcome follow a certain pattern. To allow information borrowing across levels, we propose structural model assumptions that can be placed based on prior scientific knowledge to improve the precision of the estimates. We also show sensitivity analysis to verify the structures in different settings.

14. POSTERS: PERSONALIZED MEDICINE AND DYNAMIC TREATMENT REGIMENS

14a. Survival Bandits

Yinghao Pan, University of North Carolina at Charlotte

We consider a contextual survival bandit setting, a variant of the classical multi-armed bandit problem in which the reward for each individual is a survival time subject to right censoring. First, we design a Thompson sampling algorithm that randomly allocates individuals to treatments in an adaptive manner based on Bayesian posterior distributions. Next, we propose a weighted M-estimator for constructing valid confidence regions using data collected from the Thompson sampling algorithm mentioned above. Asymptotic properties of the proposed weighted M-estimator are established by careful use of martingale theory.

14b. A Residual Life Value Function for Dynamic Treatment Regime Optimization via Q-Learning

Grace Rhodes, North Carolina State University

Clinicians and patients must make treatment decisions at a series of key decision points throughout disease progression. A dynamic treatment regime is a set of sequential decision rules that return treatment decisions based on accumulating patient information, like that commonly found in electronic medical record (EMR) data. When applied to a patient population, an optimal dynamic treatment regime leads to the most favorable outcome on average. Optimization with respect to maximizing remaining life expectancy is especially desirable for patients with life-threatening diseases such as sepsis, a complex medical condition that involves severe infections with organ dysfunction. Using the backward-induction algorithm "Q-learning," we introduce a method to estimate an optimal dynamic treatment regime that maximizes the expected value of the sum of residual life across all decision points at which the patient is alive. We illustrate the utility of the estimation procedure in simulation studies. We apply the Q-learning method to dynamically optimize the treatment regime of septic patients in the intensive care unit using EMR data from the MIMIC-III database.

14c. Interim Monitoring of Sequential Multiple Assignment Randomized Trials Using Partial Information

Cole Manschot, North Carolina State University

Sequential multiple assignment randomized trials (SMARTs) are the gold standard trial design to generate data for the evaluation of multi-stage treatment regimes. Interim monitoring allows early stopping; however, there are few methods for principled interim analysis in SMARTs. Because SMARTs involve multiple stages of treatment, a key challenge is that not all enrolled participants will have progressed through all treatment stages at the time of an interim analysis. We propose a doubly-robust estimator for the mean outcome under a given regime that gains efficiency by using partial information from enrolled participants regardless of their progression through treatment stages. Using the asymptotic distribution of this estimator, we derive associated Pocock and O'Brien-Fleming testing procedures for early stopping. In simulation experiments, the estimator controls type I error and achieves nominal power while reducing expected sample size relative to the method of Wu, Wang, and Wahed (2021). We provide an illustrative application of the proposed estimator using a case study based on a recent SMART evaluating behavioral pain interventions for breast cancer patients.

15. POSTERS: PREDICTION, DECISION MAKING, AND META-ANALYSIS

15a. Dynamic Prediction Using Cox Survival Neural Network with Time-Dependent Covariates

Lang Zeng, University of Pittsburgh

The target of dynamic prediction is to provide individualized risk predictions over time which can be updated as new data become available. Motivated by establishing a progression prediction model for a progressive eye disease, the age-related macular degeneration (AMD) using longitudinal fundus images, we propose a time-dependent Cox model-based neural network to predict disease progression over time. We evaluate and compare our proposed method with joint modeling and landmarking approaches through comprehensive simulations using two time-dependent accuracy metrics, Brier score and dynamic AUC. We apply the proposed method to a large AMD study, the Age-Related Eye Disease Study (AREDS), in which over 50,00 fundus images were captured over a period of 12 years for more than 4000 participants. The results indicate that our model achieves satisfactory prediction performance.

15b. XGBoost Model of Cholesterol-Related Genes is Prognositic in Colon Cancer

Xiuxiu Yang, Indiana University

Background: Colon cancer (CC) is one of the most common cancers with high mortality, and recent evidence has revealed that a high-cholesterol diet is related to elevated risk. However, it is still unclear how the association can predict prognosis and stratify for precision medicine. Methods: To study the connection between cholesterol and CC prognosis, transcriptomic data were obtained from TCGA and GEO. Differential expression analysis followed by LASSO was performed to select cholesterol-related genes. Then, eXtreme Gradient Boosting (XGBoost) model was trained on prognostic genes. Results: There were 1167 DEGs identified, and bile secretion pathway was enriched. Our multivariate Cox model can stratify patients from TCGA (p-value <0.01), GSE17538 (p-value=0.03) and GSE39582 (p-value=0.05). Our XGBoost model can stratify patients from TCGA (p-value <0.01), GSE17538(p-value =0.02) and GSE33113(p-value<0.01). Overall, our XGBoost model was better able to stratify patients than a traditional regression model. Our findings show that the XGBoost model can improve patient stratification and highlight prognostic potentials of cholesterol pathways in CC.

15c. Testing Stationarity in Sequential Decision Making

Jitao Wang, University of Michigan

Reinforcement Learning (RL) is a powerful technique that allows an autonomous agent to learn an optimal policy to maximize the expected return. Central to the optimality of various RL algorithms is the stationarity assumption that requires time-invariant state transition and reward functions. However, deviations from stationarity over extended periods often occur in real-world applications like robotics control, health care and digital marketing, resulting in sub-optimal policies learned under stationary assumptions. In this paper, we propose a doubly robust procedure for testing the stationarity assumption and change point detection. The proposed test controls the type-I error and has good power properties even in high dimensions. Extensive comparative simulations and a real-world interventional mobile health example demonstrate the advantages of the proposed method in change point detection and optimizing long-term rewards in high-dimensional nonstationary environments.

15d. Constrained Maximum Likelihood Approach to Subgroup Calibration

Sarah Hegarty, University of Pennsylvania

Risk prediction algorithms learned using electronic health records (EHR) data offer a powerful tool for supporting clinical decision-making. However, there are abundant challenges in repurposing EHRs for model training, including a potential lack of representativeness in the patient population, differences in information availability and noisy labels. The choice of loss function also comes with trade-offs between maximizing predictive accuracy and robustness to distributional differences between subgroups. Differences is model performance between subgroups can lead to a lack of fairness. Recent efforts to address algorithm fairness largely focus on equalizing performance metrics like false positive rate. However, enforcing error rate equality can result in losses in accuracy and calibration. Calibration is critical for clinical decision making. We conduct an empirical study of a model revision approach that aims to improve subgroup calibration. After learning an initial working model, we fit a logistic regression model using the predicted score and a subvector of features. We constrain the maximization procedure to enforce mean calibration within predefined subgroups.

15e. Use of Previously Published Results from External Studies for Improving Statistical Inference

Sergey Tarima, Medical College of Wisconsin

Typically, researchers collect and analyze their own data, or aggregate previously published results and perform meta-analysis. In our work, we make statistical inference using both researcher's own data (experimental or observational) and the data extracted from external data sources reported by an estimate (such as sample mean) and its standard error. If this information comes from study with similar data collection plan from a similar population, the external information can substantially decrease variance in the estimation of the quantity of interest, whereas external information coming from external data sources with different data collection protocols and/or different populations of patients can seriously bias statistical inference with the use of this information. We consider frequentist and Bayesian methods for using external information and compared their performance in different simulation scenarios. Our motivating example shows how association between student’s gestational age and 3rd grade standardized reading score can be improved with additional information. We discuss cons and pros of incorporating external information.

16. POSTERS: SPATIAL/TEMPORAL AND TIME-SERIES MODELING

16a. The Utility of a Bayesian Predictive Model to Forecast Neuroinvasive West Nile Virus in the United States, 2022

Maggie McCarter, University of South Carolina

Arboviruses are an emerging global health threat that are rapidly spreading as local ecologies are impacted by various adverse factors. More than 25,000 cases of West Nile Neuroinvasive Disease (WNND) have been diagnosed in the United States, cementing WNV as of public health importance. Using the Centers for Disease Control and Prevention (CDC) ArboNET WNV data from 2000–2021, and as part of the CDC WNND forecasting challenge, this study aimed to predict WNND human cases at the county level for the contiguous US states using a spatio-temporal Bayesian negative binomial regression model. The model includes environmental, climatic, host, and demographic factors. An integrated nested LaPlace approximation approach was used to fit our model. To assess model prediction accuracy, annual counts were withheld, forecasted, and compared to observed values. The validated models were then fit to the entire dataset for 2022 predictions. This proof-of-concept mathematical geospatial modelling approach has utility for national health agencies seeking to allocate funding and resources for local vector control agencies tackling WNV and other arboviral agents.

16b. Estimating the Power to Detect Spatially Varying Coefficient Effects with Conditional Autoregressive (CAR) Models: A Simulation Study Using Social Determinants of Health Screening Data

Reid DeMass, University of South Carolina

Social determinants of health (SDoH) refer to the environmental conditions in which individuals live, work and are born and are of interest to health care systems. Research has shown that both SDoH needs and their association with health care resource use vary spatially. One method to assess spatial variation is to assume a spatially varying coefficient model with a conditional autoregressive (CAR) prior on the coefficients of SDoH-need. The power of such models to detect spatial variation depends on the prevalence and spatial distribution of the SDoH variable, which is binary in structure. Pilot data from Prisma Health was used to simulate an SDoH variable and a resource use count variable. The count followed a zero-inflated negative binomial distribution and spatial coefficients were estimated using integrated nested Laplace approximation. The model was simulated 500 times for increasing SDoH prevalence by randomly selecting patients to possess the SDoH need. The aim of this simulation study was to assess whether some threshold of SDoH prevalence exists at which the model power reaches a desired level to guide spatial analysis of binary variables with low prevalence.

16c. Informative Missingness in Stochastic Volatility Models

Gehui Zhang, University of Pittsburgh

Mobile devices that collect activity and emotional data in one's natural environment are becoming an increasingly popular and powerful tool for clinicians and researchers.Volatility, which is a measure of dispersion or variation, is an interpretable and predictive metric from psychosocial and behavioral data. Existing methods for estimating stochastic volatility assume MAR, which is not consistent with the observational nature of data from mobile devices. The main focus of this talk is estimating stochastic volatility while accounting for informative missingness. We developed an imputation method based on Tukey's representation under the stochastic volatility model with a MNAR mechanism. We incorporated the novel imputation approach into a Particle Gibbs with Ancestor Sampling method to provide an efficient method for conducting inference on stochastic volatility with informed missingness. The performance of the method is illustrated through simulation studies and in the analysis of multi-modal ecological momentary assessment data from a study of suicidal ideation and behavior in young adults with a history of suicidal thoughts and behaviors.

16d. Two-Fold Random Slopes Models for the Analysis of Glaucomatous Visual Fields

Chen Zhao, University of Miami

Static automated perimetry is the most common form of measuring visual fields. In glaucomatous eyes, it is generally assumed that visual loss can be due to either a global loss or focal loss of visual sensitivities. In this work, we will consider modeling focal visual loss over time in a series of 500 glaucoma patients at the Bascom Palmer Eye Institute. Specifically, we have repeated visual field tests for each individual along with information on three different types of focal clustering patterns. The goal is to better identify fast progressors amongst this group of patients.

To model this data, we develop two-fold random slope models which can be seen as an extension of the two-fold small area estimation models of Torabi and Rao (2014). We also include spatial structure in the random effects to account for the orientation of focal clusters. Our model borrows strength across individuals and across connected focal clusters resulting in more accurate estimates (as measured by estimated MSPE) of individual rates of visual sensitivity decline. Confidence intervals are estimated using parametric bootstrapping.

16e. Estimation of Prescribed Wildfire Attributed PM2.5 Pollution Impact Across California, 2008-2017

Brandon Feng, North Carolina State University

Increased smoke pollution from the growing wildfire crisis in California poses many public health and safety dangers. Therefore, accurately modeling fine air pollution particulate matter PM2.5 is a critical task. The current approach utilizes a mathematical dispersion model to estimate PM2.5 arising from both naturally occurring wildfires and prescribed burns. This method, however, has shortcomings such as overestimating wildfire PM2.5. To address these issues, we propose a bias correction on the estimated wildfire pollution levels via a spatial two-stage partial covariate penalized ridge regression. This model approximates the total PM2.5 across various sites as a function of the dispersion model estimates, smoke status and spline functions, the last of which estimates background pollution effect. We obtain interpretable bias correction coefficients for the effect of prescribed and natural wildfires and show that days with smoke are estimated more accurately. The health burden stemming from wildfire PM2.5 is then estimated across various population groups using this fitted model.

17. POSTERS: STATISTICAL GENETICS

17a. Using Tissue-Specific Genetic Variation in Mendelian Randomization to Map the Influence of Education and Intelligence on the Risk of Alzheimer's Disease

Ryan Liu, University of Pennsylvania

The causality between tissue-mediated intelligence and Alzheimer’s Disease risk is unclear. We conducted a two-sample multivariable Mendelian randomization study to analyze how tissue-specific gene expression mediates the effect of intelligence on Alzheimer’s Disease risk. Tissue-specific expression-mediated intelligence was associated with Alzheimer’s Disease risk independently from educational attainment with an odds ratio of -1.59 (95% CI: -2.73, -0.443; p = 6.57×10-3). This study provides evidence supporting causal association between tissue-specific gene expressions and the effect of intelligence on Alzheimer’s Disease risk. The findings of this study suggest that cognitive training combined with nutrient supplements that increase brain reserve may result in reduced Alzheimer’s Disease risk. However, the methods for which cognitive training can be conducted remain unclear. We recommend future investigation into the feasibility of lowering Alzheimer's disease risk through food supplements that improve brain cognitive functions.

17b. A Comprehensive Assessment of Cell Type-specific Differential Expression Methods in Bulk Data

Guanqun Meng, Case Western Reserve University

Accounting for cell-type compositions has been very successful at analyzing high-throughput data from heterogeneous tissues. Differential gene expression analysis at cell type-level is becoming increasingly popular. Although several computational methods have been developed to identify cell type-specific differentially expressed genes (csDEG) from RNA-seq data, a systematic evaluation is yet to be performed. Here, we thoroughly benchmark eight recently published methods: CellDMC, CARseq, TOAST, LRCDE, CeDAR, TCA, csSAM and DESeq2, for a comprehensive comparison. In simulation studies, we benchmark these methods under various scenarios of baseline expression levels, sample sizes, cell-type compositions, expression level alterations, technical noises, and biological dispersions. Real data analysis on inflammatory bowel disease, lung cancer, and autism provides both gene-level and pathway-level evaluations. We find that csDEG calling is strongly affected by effect size, baseline expression level, and cell type compositions. Results imply that csDEG discovery is a challenging task, with room to improvement on handling low signal-to-noise ratio and low expression genes.

17c. iIMPACT: Integrating Image-based and Molecular Profiles to Analyze and Cluster Spatial Transcriptomics Data

Xi Jiang, Southern Methodist University

The breakthrough in spatially resolved transcriptomics (SRT) has enabled comprehensive molecular characterization with high resolution while preserving spatial information. Meanwhile, pathology image analysis powered by artificial intelligence enables the histology characterization of single cells. Understanding the spatial organization of cells and their heterogeneous gene expression profiles will provide deeper biological insights. In this study, we developed iIMPACT to cluster and analyze SRT data. It is implemented by an interpretable Bayesian finite mixture model with the Markov random field for spatial domain segmentation. It combines a Gaussian component to model the gene expression profiles, a multinomial component for cell abundances, and the spatial information by a Markov random field prior. Applying our method to publicly available datasets, we found that iIMPACT outperforms existing clustering methods in terms of spatial domain segmentation accuracy and has high biological interpretability in defining characteristics of detected domains. The downstream analysis of spatial variable genes shows that more cancer driver genes can be detected through iIMPACT.

17d. Haplotype Association Based on Recombination Disequilibrium

Hongyan Xu, Augusta University

Haplotype-based association analysis has several advantages over single-SNP association analysis. However, to date all associations between haplotypesand diseases have not excluded recombination interference among multiple SNP loci within haplotypes and hence some results might be confounded by recombination interference. A new haplotype-diseaseassociation method was developed based on recombination disequilibrium (RD). Applying this methodto a SNP haplotype dataset of 210 Alzheimer disease (AD) cases and 159 nondemented elderly controls, we successfully found that some pairs of sister-haplotypescontaining ApoEgene were associated with risk for ADunder no RD butall those without ApoE-ε3 and ApoE-ε4 were not associated with risk for the disease and sister-haplotype pairswithin genes IL-13 and COMT were notassociated with risk for breast cancer. In addition, none of sister-haplotype pairs in IL-17A gene wasdetected to be associated with risk for coronary artery disease. All the previously reported associations of haplotypes within thesegenes with risk for these diseasesmight bedue to strong RD and/or inappropriate haplotype pairs

17e. Methods for Multi-Phenotype Genetic Colocalization Analysis in a Single Cohort

Sarah Hanks, University of Michigan

Genetic colocalization is the process of identifying genetic variants that are causal for multiple association signals at a single locus. Most existing methods explicitly assume that the traits are measured in independent, non-overlapping samples. However, there are cohorts with multiple traits collected on the same samples, such as the METSIM study measuring ~1.4K metabolites in ~6K individuals. To investigate the consequences of violating the independent cohort assumption, we used genetic data from the METSIM study to simulate two continuous traits, varying the proportion of the error variance that is attributed to individual-level confounding. We used fastENLOC to perform colocalization on the simulated data. We found that as the individual-level confounding increased, the false positive rate increased, although type I error rates remained well controlled due to the low power of colocalization methods. We propose a new method to perform colocalization in a single cohort by explicitly estimating the shared individual-level confounding across the two traits. In the future, we will use this method to perform colocalization analysis with metabolite data in the METSIM study.

17f. Efficient and Accurate Frailty Model Approach for Genome-Wide Survival Association Analysis in Large-Scale Biobanks

Rounak Dey, Harvard T.H. Chan School of Public Health

With decades of electronic health records linked to genetic data, large biobanks provide unprecedented opportunities for systematically understanding the genetics of the natural history of complex diseases. Genome-wide survival association analysis can identify genetic variants associated with ages of onset, disease progression and lifespan. We propose an efficient and accurate frailty model approach for genome-wide survival association analysis of censored time-to-event phenotypes by accounting for both population structure and relatedness. Our method utilizes state-of-the-art optimization strategies to reduce the computational cost. The saddlepoint approximation is used to allow for analysis of heavily censored phenotypes (>90%) and low frequency variants (down to minor allele count 20). We demonstrate the performance of our method through extensive simulation studies and analysis of five time-to-event phenotypes, including lifespan, with heavy censoring rates (90.9% to 99.8%) on ~400,000 UK Biobank participants with white British ancestry and ~180,000 individuals in FinnGen. We further analyzed 871 TTE phenotypes in the UK Biobank.

WITHDRAWN 17g. On Asymptotic Distributions of Several Test Statistics for Familial Relatedness in Linear Mixed Models

Tao Wang, Medical College of Wisconsin

In this study, the asymptotic distributions of the likelihood ratio test (LRT), the restricted likelihood ratio test (RLRT), the F and the Sequence Kernel Association Test (SKAT) statistics for testing an additive effect of the expected familial relatedness (FR) in a linear mixed model are examined based on an eigenvalue approach. The covariance structure for modeling the FR effect in a LMM is illustrated first. Then, the multiplicity of eigenvalues for the log-likelihood and restricted log-likelihood is constructed under a replicate family setting (RFS) and extended to a more general replicate family setting (GRFS) as well. After that, the asymptotic null distributions of LRT, RLRT, F and SKAT statistics under GRFS are derived. The asymptotic null distribution of SKAT for testing genetic rare variants is also established. In addition, a simple formula for sample size calculation is provided based on the restricted maximum likelihood estimate of the effect size for the expected FR. A power comparison of these test statistics on hypothesis test of the expected FR effect is made via simulation. The four test statistics are also applied to a data set from the UK Biobank.

17h. Winner's Curse Lifted Robust Mendelian Randomization with Summary Data

Zhongming Xie, University of California, Berkeley

In the past decade, the increased availability of genome-wide association studies (GWAS) summary data has popularized Mendelian Randomization (MR) to conduct causal inference. MR analyses are robust to reverse causation bias and the presence of unmeasured confounders by incorporating single nucleotide polymorphisms (SNPs) as genetic instrumental variables (IV). Nevertheless, classical MR analyses with summary data may still produce biased estimates due to winner's curse and genetic pleiotropy. To address these two issues and establish valid causal conclusions, we propose a unified MR framework with summary data, which removes the winner's curse and eliminates invalid IVs with pleiotropy. Different from existing robust MR literature, our framework does not require the pleiotropy effects to follow any parametric distribution. To further reduce the randomness of IV screening, we adopt bagging to construct our final causal effect estimator. Under appropriate conditions, we show our proposed estimator is asymptotically normal and its variance can be well estimated. We illustrate the performance of our proposed estimator through Monte Carlo simulations and case studies.

Monday, March 20, 2023 | 8:30-10:15 AM

1. NEW INSIGHTS ON STATISTICAL MODELING FOR BRAIN CONNECTIVITY AND BRAIN IMAGING GENOMICS

Organizer: Yize Zhao, Department of Biostatistics, Yale University
Chair: Zhe Sun, Department of Biostatistics, Yale University

8:30-8:55 AM

Mapping the Causal Genetic Imaging Clinical (CGIC) Pathway of Brain-Related Disorders

Hongtu Zhu, University of North Carolina at Chapel Hill

We develop a statistical learning framework for mapping the causal genetic imaging clinical (CGIC) pathway motivated by the joint analysis of genetic, imaging, and clinical (GIC) data collected in many large-scale biomedical studies, such as the UK Biobank study and the Alzheimer's Disease Neuroimaging Initiative (ADNI) study. We develop a joint model selection and estimation procedure by embedding imaging data in the reproducing kernel Hilbert space and imposing the penalty for the coefficients of scalar variables. We systematically investigate the theoretical properties of scalar and functional efficient estimators, including non-asymptotic error bound, minimax error bound, and asymptotic normality. We apply the proposed method to multiple imaging genetic datasets to identify important features from several millions of genetic polymorphisms and study the effects of a certain set of informative genetic variants and the hippocampus surface on thirteen cognitive variables.

8:55-9:20 AM

Tumor Radiogenomics in Gliomas with Bayesian Layered Variable Selection

Veerabhadran Baladandayuthapani, University of Michigan

We propose a statistical framework to integrate radiological magnetic resonance imaging (MRI) and genomic data to identify the underlying radiogenomic associations in lower grade gliomas (LGG). We devise a novel imaging phenotype by dividing the tumor region into concentric spherical layers that mimics the tumor evolution process. MRI data within each layer is represented by voxel-intensity-based probability density functions which capture the complete information about tumor heterogeneity. Under a Riemannian-geometric framework these densities are mapped to a vector of principal component scores which act as imaging phenotypes. Subsequently, we build Bayesian variable selection models for each layer with the imaging phenotypes as the response and the genomic markers as predictors. Our novel hierarchical prior formulation incorporates the interior-to-exterior structure of the layers, and the correlation between the genomic markers. In the LGG context, genes implicated with survival and oncogenesis are identified, which could potentially serve as early-stage diagnostic markers for disease monitoring, prior to routine invasive approaches.

9:20-9:45 AM

Omics-Imaging Data Integration via Mediation Analysis with High-Dimensional Exposures and Mediators

Yi Zhao, Indiana University

Motivated by an imaging proteomics study for Alzheimer’s disease (AD), in this article, we propose a mediation analysis approach with high-dimensional exposures and high-dimensional mediators to integrate data collected from multiple platforms. The proposed method combines principal component analysis with penalized least squares estimation for a set of linear structural equation models. The former reduces the dimensionality and produces uncorrelated linear combinations of the exposure variables, whereas the latter achieves simultaneous path selection and effect estimation while allowing the mediators to be correlated. Applying the method to the AD data identifies numerous interesting protein peptides, brain regions, and protein-structure-memory paths, which are in accordance with and also supplement existing findings of AD research. Additional simulations further demonstrate the effective empirical performance of the method.

9:45-10:10 AM

A Predictor-Informed Bayesian Approach for Dynamic Functional Connectivity

Michele Guindani, University of California, Los Angeles

Time-Varying Functional Connectivity investigates how the interactions among brain regions vary over the course of an fMRI experiment. The transitions between different individual connectivity states can be modulated by changes in underlying physiological mechanisms that drive functional network dynamics, e.g., changes in attention or cognitive effort as measured by concurrent measurements. In this talk, I will describe Bayesian approaches for estimating dynamic functional networks as a function of time-varying exogenous physiological covariates that are simultaneously recorded in subjects during the fMRI experiment. We apply our modeling framework on resting-state experiments where fMRI data have been collected concurrently with pupillometry measurements, leading us to assess the heterogeneity of the effects of changes in pupil dilation on the subjects' propensity to change connectivity states.

2. STATISTICAL AND MACHINE LEARNING METHODS FOR CAUSAL INFERENCE WITH SPATIAL DATA

Organizer/Chair: Corwin Zigler, University of Texas as Austin

8:30-8:55 AM

Causal Inference for Spatial Environmental Exposures

Elizabeth Ogburn, Johns Hopkins Bloomberg School of Public Health

Spatial confounding is well known to be a challenge for the study of environmental exposures, but many existing methods do not address the issue of confounding as it is understood in the field of causal inference. Furthermore, they can result in misleading effect estimates when the true effects are heterogeneous or nonlinear. In this work we take a causal inference perspective to defining spatial confounding, describe the relationship between spatial confounding and spatial dependence, and advocate the use of more flexible models, specifically double machine learning (DML) methods, for estimating causal effects of environmental exposures in the presence of spatial confounding. We demonstrate the advantages of the DML approach analytically and via extensive simulation studies, and we apply the method to study the link between birthweight and air pollution exposure in the state of California. This is joint work with Brian Gilbert, Abhirup Datta, and Joan Casey.

8:55-9:20 AM

Spatial Causal Inference with Preferential Sampling

Brian Reich, North Carolina State University, University of Texas at Austin

Environmental data are often observational and spatially dependent, making casual treatment effects difficult to estimate. Unmeasured spatial confounders, i.e., spatial variables correlated with both the treatment and response, can induce bias and invalidate inference. Spatial data can also be subject to preferential sampling, where the locations of the samples are driven by unmeasured covariates or even the assumed value of the response of interest. We propose a method that simultaneously accounts for unmeasured confounders in both the sampling locations and treatment allocation. We prove that the key parameters in the model are identifiable and show via simulations that the causal effect of interest can be reliably estimated under the assumed model. The proposed method is applied to study the effect of marine protection areas on fish biodiversity.

9:20-9:45 AM

Weather2vec: Representation Learning for Causal Inference with Non-Local Confounding in Air Pollution and Climate Studies

Mauricio Tec, Harvard University

Non-local confounding (NLC) is an often overlooked source of bias in causal effect estimation with spatial data, occurring when the treatments and outcomes of a given unit are affected by covariates of other (perhaps nearby) locations. In this talk, we first formalize NLC using the potential outcomes framework, discussing its connection with the related but distinct phenomenon of interference. Next, we present weather2vec, a broadly applicable framework using balancing scores theory and convolutional neural networks, specifically a U-net, to learn representations of NLC for each observational unit that one can use to adjust for confounding in conjunction with traditional causal inference methods. Weather2vec is motivated by environmental scientific studies wherein meteorology is a known confounder that simple ad hoc functions cannot account for. We demonstrate the framework in two real applications. The first one measures the effect of an intervention to reduce emissions of air pollutants. The second one aims to deconvolve climate variations from policy factors when characterizing long-term trends of air pollution exposure.

9:45-10:10 AM

Spatial Causal Inference in the Presence of Unmeasured Confounding and Interference

Georgia Papadogeorgou, University of Florida

Causal inference in spatial settings is met with unique challenges and opportunities. On one hand, a unit's outcome can be affected by the exposure at many locations, leading to interference. On the other hand, unmeasured spatial variables can confound the effect of interest. We illustrate that spatial confounding and interference can manifest as each other, meaning that investigating the presence of one can lead to wrongful conclusions in the presence of the other. To address this, we propose an approach to account for unmeasured spatial confounding and interference simultaneously. Our approach is based on simultaneous modeling of the exposure and the outcome while accounting for the presence of spatially-structured unmeasured predictors of both variables.

3. FRONTIERS IN ANALYSIS OF MICROBIOME DATA, FROM CONCEPTS, METHODS TO APPLICATION

Organizer/Chair: Yijuan Hu, Emory University

8:30-8:55 AM

What Can We Learn about the Bias of Microbiome Studies from Analyzing Data from Model Communities?

Glen Satten, Emory University School of Medicine

Data from both 16S and shotgun metagenomics studies are subject to biases that cause the observed relative abundances of taxa to differ from their true values. Model community analyses, in which the relative abundances of all taxa in the sample are known by construction, offer hope that these biases can be measured. However, it is unclear whether the bias we measure in a mock community is the same as measured samples with spiked-in taxa having known relative abundance, or if the bias in spike-in samples is the same as the bias in real (e.g., biological) samples. We consider these questions in the context of 16S rRNA measurements on three sample sets: the commercially-available Zymo cells model community; the Zymo model community mixed with Swedish Snus, a smokeless tobacco product that is virtually bacteria-free; and a set of commercially-available smokeless tobacco products. Each sample set was subject to four different extraction protocols. We determine whether the patterns of bias observed in each set of samples are the same; i.e., can we learn about the bias in the commercially-available smokeless tobacco products by studying the Zymo cells model community.

8:55-9:20 AM

Genome-Microbiome Association Testing via Integrated Zero-Inflated Quantile Processes

Wodan Ling, Weill Cornell Medicine

Associations between genome and microbiome are largely unknown yet beneficial for medical practice. For example, understanding how host genetic variations shape the vaginal microbiome to affect preterm births helps clinicians treat those at higher risk of delivering preterm in advance. Most existing analyses focus on the association between SNP(s) and the conditional mean of individual taxon abundance. However, these approaches are limited due to the highly zero-inflated and dispersed nature of microbiome data, let alone the genome-microbiome association may be heterogeneous. We propose an integrated zero-inflated quantile process approach. It uses a test in logistic regression to accommodate the zero counts of microbiome data and integrates quantiles of the non-zero abundance over a process while adjusting for zero inflation. Our approach is robust to the irregular microbiome distributions and is powerful regardless of the locations of association signals. Simulation and real data application show that the proposed approach has equivalent or higher power than existing ones while controlling type I error well.

9:20-9:45 AM

Association of Gut Microbiota with HIV infection and AIDS

Shyamal Peddada, Biostatistics and Computational Biology Branch, NIEHS

There is a growing evidence in the literature demonstrating associations between gut microbiota and inflammation and immune response. Using patient stool and blood samples preserved from the beginning of the HIV/AIDS pandemic in the 1980s, we discovered microbes that are differentially abundant months before patients seroconverted, and years before HIV positive patients developed AIDS. Specifically, reduced abundance of several species of commensal gut bacteria such as Bacteroides, Alistipes, Akkermansia, Ruminococcus, promotors of T-regulatory cell function, and an increase in the abundance of Prevotella, a proinflammatory taxa, were observed in men months before becoming HIV positive. Among HIV positive men, there was an increased abundance of Prevotella and several species of the family Lachnospiraceae, and a reduction in commensal bacteria such as Blautia among men who rapidly developed AIDS compared to those who took longer than 10 years. A positive association between propionic acid, a short chain fatty acid, and the CD4/CD8, and an increase in cytokines such as IL-6, sCD14, and sCD163, was also present months before a patient was HIV positive.

9:45-10:10 AM

Multi-Scale Adaptive Differential Abundance Analysis in Microbial Compositional Data

Shulei Wang, University of Illinois at Urbana-Champaign

Differential abundance analysis is an essential and commonly used tool to characterize the difference between microbial communities. However, identifying differentially abundant microbes remains a challenging problem because the observed microbiome data is inherently compositional, excessive sparse, and distorted by experimental bias. Besides these major challenges, the results of differential abundance analysis also depend largely on the choice of analysis unit, adding another practical complexity to this already complicated problem. In this work, we introduce a new differential abundance test called the MsRDB test, which embeds the sequences into a metric space and integrates a multi-scale adaptive strategy for utilizing spatial structure to identify differentially abundant microbes. Compared with existing methods, the MsRDB test can detect differentially abundant microbes at the finest resolution offered by data and provide adequate detection power while being robust to zero counts, compositional effect, and experimental bias in the microbial compositional data set.

4. RECENT ADVANCES IN METHODS OF ANALYSIS FOR MISMEASURED OR PREDICTED OUTCOMES IN BIOMEDICAL RESEARCH

Organizer/Chair: Pamela Shaw, Kaiser Permanente Washington Health Research Institute

8:30-8:55 AM

Generalized Network Structured Models with Mixed Responses Subject to Measurement Error and Misclassification

Grace Yi, University of Western Ontario

We consider the problem with mixed binary and continuous responses that are subject to mismeasurement and associated with complex structured covariates. We start with the case where data are precisely measured. We propose a generalized network structured model and develop a two-step inferential procedure. Furthermore, we extend the development to accommodating mismeasured responses. We consider two cases where the information on mismeasurement is either known or estimated from a validation sample. Theoretical results are established and numerical studies are conducted to evaluate the finite sample performance of the proposed methods.

8:55-9:20 AM

Accurate Validation of Clinical Prediction Models in the Presence of Outcome Misclassification

Rebecca Coley, Kaiser Permanente Washington Health Research Institute

Clinical prediction models estimated and validated using routinely collected clinical data are growing in popularity, but mismeasurement and misclassification are common in phenotypes derived from the electronic health record (EHR). This research focuses on misclassification of a rate-event outcome in a prediction study. We consider a setting where "gold standard" outcome measurements are ascertained in a subsample of the data available for prediction modeling via expert chart review and compared to EHR-derived phenotypes to estimate misclassification rates. We present a method to adjust for non-differential and differential outcome misclassification when evaluating the performance of a prediction model. We demonstrate statistical properties of our method in plasmode simulation studies in which simulated data are based on real-world clinical data on several million outpatient mental health visits and (possibly misclassified) suicide outcomes.

9:20-9:45 AM

Model Calibration and Evaluation via Optimal Sampling in Electronic Health Record Data

Jinbo Chen, University of Pennsylvania

One major challenge in using electronic health record (EHR) data for research is that patients' phenotypes need to be inferred from observed data elements. We study efficient methods for validating phenotyping models in a new EHR system. Considering that phenotyping is performed only on a subsample of patients selected based on the model that is being validated, our method for estimating accuracy metrics gained efficiency by efficiently utilizing data from both selected and un-selected patients. We further study subsampling procedure for selective phenotyping of a subgroup of patients towards more efficient estimation of accuracy metrics. Results from extensive simulation studies supported the efficiency of our proposed methods.

9:45-10:10 AM

Multi-Wave Validation Sampling to Improve Estimates Derived from Electronic Health Record Data

Bryan Shepherd, Vanderbilt University Medical Center

Researchers were interested in measuring the association between maternal weight gain during pregnancy and the risks of childhood obesity and asthma using data derived from the electronic health record (N=10,335 mother-child pairs). Because of data quality concerns, we validated all analysis variables (outcomes, exposures, and covariates; 19 total variables) for 996 mother-child pairs. We a priori decided to combine the validated (phase-2) and unvalidated (phase-1) data in our analyses using generalized raking / augmented inverse probability weighting techniques. The optimal validation sample for such an analysis depends on quantities that are generally not known until validation data are collected. Therefore, we designed and carried out multiple waves of validation, where earlier waves of data validation were used to inform the optimal sample for selecting future waves. In this talk I will present our multi-wave validation strategies, study findings, lessons learned, and additional areas of research motivated by our experience.

5. RECENT ADVANCES IN SURVIVAL ANALYSIS FOR DATA INTEGRATION FROM DIVERSE SOURCES

Organizer/Chair: Kevin He, University of Michigan

8:30-8:55 AM

Scalable Data Integration in Genome-wide Association Studies (GWAS) through Generalized Method of Moments

Nilanjan Chatterjee, Johns Hopkins University

We have shown earlier (Kundu et al., Biometrika, 2019) that generalized method of moments (GMM) can provide a unified framework for building complex models through integration of information across multiple disparate data sources. In this work, we consider data integration in the context of large-scale genetic association studies which requires parallel analysis of hundreds of thousands or even millions of genetic variants. We now develop score-tests and one-step optimization techniques within the GMM framework for scalable hypothesis testing and fast parameter estimation for the analysis of GWAS scale datasets. We will show how this general framework can be useful for data integration across individual level data and external GWAS summary-statistics to carry out a whole set of cutting-edge applications which requires incorporation of non-genetic covariate into the underlying models for data analysis.

8:55-9:20 AM

Hierarchical Heterogeneity Analysis for Cancer Survival Based on Pathological Imaging Features

Shuangge Ma, Yale University

In cancer research, supervised heterogeneity analysis has important implications. Recently, pathological imaging features, which are generated as a byproduct of biopsy, have been shown as effective for modeling cancer survival (and other outcomes), and a handful of supervised heterogeneity analysis has been conducted based on such features. There are two types of pathological imaging features, which are extracted based on specific biological knowledge and using automated imaging processing software, respectively. Using both types of pathological imaging features, our goal is to conduct supervised cancer heterogeneity analysis that satisfies a hierarchical structure. More specifically, the first type of imaging features defines a rough structure, and the second type defines a nested and more refined structure. A penalization approach is developed, which is motivated by penalized fusion and sparse group penalization. The analysis of overall survival of lung adenocarcinoma patients further demonstrates the practical utility of this novel approach.

9:20-9:45 AM

Combining Primary Cohort Data with External Aggregate Information without Assuming Comparability

Jing Ning, The University of Texas MD Anderson Cancer Center

In comparative effectiveness research (CER) for rare types of cancer, it is appealing to combine primary cohort data containing detailed tumor profiles together with aggregate information derived from cancer registry databases. Such integration of data may improve statistical efficiency in CER. A major challenge in combining information from different resources, however, is that the aggregate information from the cancer registry databases could be incomparable with the primary cohort data, which are often collected from a single cancer center or a clinical trial. We develop an adaptive estimation procedure, which uses the combined information to determine the degree of information borrowing from the aggregate data of the external resource. We establish the asymptotic properties of the estimators and evaluate the finite sample performance via simulation studies. The proposed method yields a substantial gain in statistical efficiency over the conventional method using the primary cohort only, and avoids undesirable biases when the given external information is incomparable to the primary cohort.

9:45-10:10 AM

Semiparametric Estimation of the Transformation Model by Leveraging External Aggregate Data in the Presence of Population Heterogeneity

Chiung-Yu Huang, University of California at San Francisco

Leveraging information in aggregate data from external sources to improve estimation efficiency and prediction accuracy with smaller-scale studies has drawn much attention in recent years. Yet, conventional methods often either ignore uncertainty in the external information or fail to account for the heterogeneity between internal and external studies. This article proposes an empirical likelihood-based framework to improve the estimation of the semiparametric transformation models by incorporating information about the t-year subgroup survival probability from external sources. The proposed estimation procedure incorporates an additional likelihood component to account for uncertainty in the external information and employs a density ratio model to characterize population heterogeneity. We establish the consistency and asymptotic normality of the proposed estimator and show that it is more efficient than the conventional pseudo-partial likelihood estimator without combining information. The proposed methodologies are illustrated with an analysis of a pancreatic cancer study.

6. RECENT ADVANCEMENT OF ADAPTIVE DESIGN IN THE ERA OF PRECISION MEDICINE

Organizer: Jingshen Wang, UC Berkeley
Chair: Feifang Hu, George Washington University

8:30-8:55 AM

A Bayesian Decision-Theoretic Framework to Adaptively Estimate Personalized Minimum Effective Combinations of Sedentary Breaks to Reduce Cardiometabolic Risks

Ken Cheung, Columbia University

We address the problem of estimating personalized minimum effective combinations of multi-dimensional treatments. For context, we will describe a behavioral intervention study where we randomize sedentary breaks to participants in a crossover fashion with an objective to reduce their glucose and/or blood pressure under controlled environment. Each sedentary break regimen is defined by two elements: break frequency and duration. The trial aims to identify minimum combinations of frequency and duration that shift these cardiometabolic parameters. We will describe an adaptive design (AD) based on Bayesian decision-theoretic framework motivated by this study. Briefly, the method continuously updates the target combinations and enroll new participants based on these updates using state-of-the-art AD techniques (adaptive randomization and epsilon-tapering) and constrained estimation for Bayesian hierarchical models. We will discuss improvements due to AD in terms of false discovery rate and true positive rate relative to non-adaptive balanced randomization, and how the adaptive system addresses ethical concerns in terms of maximizing benefits of trial participants

8:55-9:20 AM

Covariate Adjustment in Group Sequential and Adaptive Designs to Improve Randomized Trial Efficiency

Kelly Van Lancker, Ghent University

In clinical trials, there is potential to improve precision and reduce the required sample size by appropriately adjusting for baseline variables in the statistical analysis. This is called covariate adjustment. In practice, many covariate adjusted estimators are incompatible with group sequential and adaptive designs. This is an obstacle for realizing precision gains from covariate adjustment as these designs are commonly used for efficiency and ethical reasons. In this talk, we propose a new statistical method that orthogonalized the original (covariate adjusted) estimator so that it becomes compatible with group sequential and adaptive designs, while simultaneously increasing or leaving unchanged the estimator's precision at each analysis. Our approach allows the use of any asymptotically linear estimator, which covers many estimators used in randomized trials. Such a method is needed in order to fully leverage prognostic baseline variables to speed up clinical trials without sacrificing validity or power. We evaluate estimator performance in simulations that mimic features of a completed trial.

9:20-9:45 AM

Bayesian Predictive Platform Design for Proof of Concept and Dose Finding

Ruitao Lin, The University of Texas MD Anderson Cancer Center

Evaluating long-term benefits of potential new treatments for chronic diseases can be very time-consuming and costly. We propose a Bayesian predictive platform design that provides a unified framework for evaluating multiple investigational agents in a multistage, randomized controlled trial. The design expedites the drug evaluation process and reduces development costs by including dose finding, futility and superiority monitoring, and enrichment, while avoiding over-allocating patients to a shared placebo or active control arm. To facilitate making real-time interim group sequential decisions, unobserved long-term responses are treated as missing values and imputed from longitudional biomarker measurements. Design parameters as well as the maximum sample size are calibrated to obtain good frequentist properties. The proposed design is illustrated by a trial of three targeted agents for systemic lupus erythematosus, evaluated by their 24-week response rates. Extensive simulations show that the proposed design compares favorably to several conventional platform designs.

9:45-10:10 AM

Adaptive Experiments Toward Learning Treatment Effect Heterogeneity

Jingshen Wang, UC Berkeley

Understanding treatment effect heterogeneity has become an increasingly important task in many scientific fields. While much of the existing work in this research area has concentrated on either analyzing observational data based on untestable causal assumptions or conducting post hoc analyses of existing randomized controlled trial data, little work has gone into designing randomized experiments specifically for uncovering treatment effect heterogeneity. We develop an adaptive experimental design framework towards learning treatment effect heterogeneity. Our design framework leverages a characterization of the probability of correctly selecting the best-performing subgroup from large deviation principles, leading to a dynamic optimization problem that can be efficiently solved. Furthermore, our design unifies commonly adopted experimental strategies, including adaptive subgroup enrichment design and response adaptive design with adaptive treatment allocation. Through our theoretical and numerical investigations, we illustrate the trade-offs between complete randomization, covariate adaptive design, and our design.

7. IN REMEMBRANCE AND HONOR OF DR. EDMUND GEHAN – A PIONEER BIOSTATISTICIAN, TRAIL BLAZER, AND MENTOR

Organizers/Chairs: J. Jack Lee, MD Anderson Cancer Center; Ming Tan, Georgetown University

8:30-10:15 AM

Dr. Edmund Gehan passed away in September 2021 at the age of 92. Ed was a pioneer biostatistician, a trailblazer in cancer biostatistics, and a great mentor. He served as President of ENAR in the early 1990's. Ed's contribution to biostatistics is legendary. For over 60 years, Ed's career spanned at National Cancer Institute, Birkbeck College in London, University of Texas MD Anderson Cancer Center, University of Paris, and Georgetown University/Lombardi Comprehensive Cancer Center. Along the way, he served as the group statistician of the Southwest Oncology Group and the Intergroup Rhabdomyosarcoma Study Group. Ed loved his role in advancing oncology to benefit all cancer patients. Among many, his most significant statistical accomplishments include developing the Gehan-Wilcoxon test for survival data and the two-stage Gehan's design, which was the golden rule for Phase II trials for decades. Ed was an inspiring mentor and colleague. He hired and mentored many of today's cancer biostatisticians. Ed is a role model for younger biostatisticians. As a gifted storyteller, he wrote: "Memoir of a Number Doctor" (https://dl.bookfunnel.com/opcnutbur8). This session will benefit younger biostatisticians to learn from Ed and continue his legacy in developing novel statistical methods, providing impactful collaborations, and promoting biostatistics in research, education, and mentoring.

8. CONTRIBUTED PAPERS: ADAPTIVE DESIGN/ADAPTIVE RANDOMIZATION IN CLINICAL TRIALS

Chair: Bryan Blette, University of Pennsylvania

8:30-8:45 AM

Adjusted Critical Boundaries Under Fractional Brownian Motion

Peng Zhang, CIMS Global LLC

Classical and adaptive group sequential designs have been developed to allow a clinical trial to claim efficacy earlier while controlling the overall type I error rate. The design methodologies are based on the standard Brownian motion (Bm) model. However, violations to the standard Bm assumption may appear in some trials, such as possible dependence may exist among the observations, or between survival time and censoring. Under this circumstance, a more inclusive model, fractional Brownian motion (fBm), can be used to describe such dependence by including the Hurst exponent. In this paper, we discuss the type I error rate inflation and deflation under different scenarios if adjustment is not made properly, and introduce calculation of adjusted critical boundaries based on known or estimated Hurst exponent at an interim or at the final analysis. Moreover, the new critical boundaries can be calculated for an adaptive plan, i.e., for circumstances where future interim analyses are either added or canceled before the final analysis. Examples of alpha-spending functions under fBm, including O?Brien-Fleming type and Pocock type boundaries are discussed with simulation.

8:45-9:00 AM

Benefits of Interim Analysis Covariate Adjustment in Bayesian Group Sequential Designs

James Willard, McGill University

In conventionally randomized controlled trials, adjustment for baseline values of covariates known to be associated with the outcome (?covariate adjustment?) increases the power of the trial. Recent work has shown similar results hold for more flexible frequentist designs, such as information adaptive and adaptive multi-arm designs. However, covariate adjustment has not been characterized within more flexible Bayesian designs, despite their growing popularity. We focus on a subclass of these, Bayesian group sequential designs, which allow for early stopping at an interim analysis given evidence of treatment superiority. For these designs, we perform a simulation study to assess the impact of interim analysis covariate adjustment using a variety of adjustment models. We consider trials with several maximum sample sizes and outcome types (continuous, binary, time-to-event), as well as a real-world COVID-19 trial with a binary endpoint. It is shown that interim analysis covariate adjustment increases power and the probability of stopping the trials early, and decreases the expected sample sizes as compared to unadjusted analyses.

9:00-9:15 AM

Group Response-Adaptive Randomization in Clinical Trials

Guannan Zhai, The George Washington University

The response-adaptive randomization (RAR) procedure has been proposed and studied extensively in the literature. However, almost all procedures are based on one crucial assumption: updating the randomization after each response, which is unrealistic in many clinical trials. In this talk, we propose a new family of response adaptive randomization procedures that are updated after a group response or at a fixed time (weekly or biweekly). We show that the proposed design keeps the important theoretical properties of usually doubly adaptive biased coin design (DBCD). Numerical studies also demonstrate that the group doubly adaptive biased coin design has similar properties as the usual DBCDs under different situations. Besides, we apply the new design to a real clinical trial, illustrating the design?s advantages and practicability. This paper opens the door to studying the properties of other types of group response adaptive designs (for example, urn models). More important, the results of this paper make it easy to apply response-adaptive randomized clinical trials in practice.

9:15-9:30 AM

Treat Now: Statistical Challenges and Solutions for a Multi-Site Outpatient COVID-19 Randomized Trial

Alexander Kaizer, University of Colorado Anschutz Medical Campus

The COVID-19 pandemic led to numerous statistical and trial design challenges that needed to be addressed in a relatively short time period in order to identify potential therapeutics and treatments. One such trial was the Trial of Early Therapies During Non-hospitalized Outpatient Window for COVID-19 (TREAT NOW), which enrolled participants across the United States from multiple sites over multiple variants of the pandemic [NCT04372628]. Challenges included the design of the study as a flexible master protocol platform trial, the choice of the primary outcome as a longitudinal ordinal outcome, the decision between frequentist and Bayesian approaches, and how to implement an interim analysis for novel statistical models. Multiple solutions were considered for each challenge, and we present the decisions that ultimately led to a final two-arm comparison utilizing Bayesian methods with upstrapping for interim monitoring.

9:30-9:45 AM

Statistical Inference of Covariate-Adaptive Randomized Vaccine Clinical Trials

Fengyu Zhao, George Washington University

Covariate-adaptive randomization(CAR) procedures are frequently used in vaccine clinical trials. The evolution of vaccine efficacy is a challenging aspect of the analysis. Firstly,the risk of infection is generally very low. Secondly, the duration of protection afforded by vaccine is unclear and vaccine efficacy usually declines over time. Furthermore, when CAR procedures are used in vaccine trials, the validity of classical statistical methods is not guaranteed since they use covariate information in the process of patients' allocation. In this presentation, we will introduce some statistical methodology for the evaluation of vaccine efficacy. Specifically, we use logistic regression model (binary outcome) and Cox model(time-to-event outcome) to estimate the vaccine efficacy under complete randomization and CAR procedures. We obtain the asymptotic distribution of the test statistic under the null hypothesis. We also propose an adjustment method to achieve a valid type I error based on the asymptotic results. Numerical studies also confirm the theoretical findings and demonstrate the effectiveness of the proposed adjustment method.

9:45-10:00 AM

Selecting an Appropriate Randomization Procedure for a Randomized Controlled Trial: A Statistician's Perspective

Oleksandr Sverdlov, Novartis

The randomized controlled trial (RCT) is an established gold standard design for evidence-based medicine. Various kinds of randomization procedures are available, and choosing an “optimal” procedure for a given trial is not straightforward. In this talk we will present a systematic roadmap for the choice and application of a restricted randomization procedure in a randomized, comparative, parallel group clinical trial with equal (1:1) allocation. Important statistical considerations will be highlighted, including a tradeoff between treatment balance and allocation randomness, type I error, and power of a statistical test. Some real-life clinical trial examples will be presented to illustrate the thinking process for selecting a randomization procedure for implementation in practice. We will argue that randomization-based tests are robust and valid alternatives to likelihood-based tests and should be considered more frequently by clinical investigators.

10:00-10:15 AM

Covariate-Adaptive Randomization Procedures with Network Effects

Jialu Wang, George Washington University

Covariate-adaptive randomization procedures have been extensively studied and applied in clinical trials. In some clinical trials, patients may link together through a network. This phenomenon is usually known as network interactions. Therefore, one should incorporate both the covariates and the network information in a carefully designed randomized clinical trial to improve the estimation of the average treatment effect (ATE) for hypothesis testing in network data. In this talk, we propose a new adaptive design to balance both the network and the covariates in clinical trials and study its properties. We demonstrate the relationships between the improved balance with respect to the covariate and network and the increased efficiency for estimating the ATE in terms of the reduction of the MSE. Numerical studies are performed to evaluate the finite sample properties of the proposed procedure. The results demonstrate the advanced performance of the proposed procedure in terms of the greater comparability of the treatment groups as well as the reduction of the bias and variance for estimating the ATE under various scenarios.

9. CONTRIBUTED PAPERS: BAYESIAN LEARNING, PRIOR ELICITATION, AND PREDICTION

Chair: Mengbing Li, University of Michigan

8:30-8:45 AM

Optimal Priors for the Discounting Parameter of the Normalized Power Prior

Yueqi Shen*, University of North Carolina at Chapel Hill

The power prior is a popular class of informative priors for incorporating information from historical data in a variety of applications. When the discounting parameter is modeled as random, the normalized power prior (NPP) is recommended. The NPP defines a conditional prior for the parameters of interest given the discounting parameter and a marginal prior for the discounting parameter. In this work, we explore the construction of optimal priors for the discounting parameter in an NPP. In particular, we are interested in achieving the dual objectives of encouraging borrowing when the historical and current data are compatible and limiting borrowing when they are in conflict. We propose intuitive procedures for eliciting the shape parameters of a beta prior for the discounting parameter based on two criteria, the Kullback-Leibler divergence and the Bayesian Mean Squared Error. In addition, we prove that the marginal posterior for the discounting parameter for generalized linear models converges to a point mass at zero if there is any discrepancy between the historical and current data, and that it does not converge to a point mass at one when they are fully compatible.

8:45-9:00 AM

Bayesian Learning of COVID-19 Vaccine Safety while Incorporating Adverse Events Ontology

Bangyao Zhao, University of Michigan, Ann Arbor

While vaccines are crucial to end the COVID-19 pandemic, public confidence in vaccine safety has always been vulnerable. Many statistical methods have been applied to VAERS (Vaccine Adverse Event Reporting System) database to study the safety of COVID-19 vaccines. However, none of these methods considered the adverse event (AE) ontology although AEs are naturally related. Explicitly bringing AE relationships into the model can aid in detecting true AE signals amid the noise while reducing false positives. We propose a Bayesian graph-assisted signal selection (BGrass) model to evaluate the risk of all AEs simultaneously while incorporating the dependence network between AEs under a logistic regression framework. We also propose a negative control approach to mitigating the reporting bias and an enrichment approach to detecting AE groups of concern. for posterior computation, we construct a novel equivalent model representation and develop an efficient Gibbs sampler using P?lya-gamma data augmentation. The performance of BGrass is demonstrated via extensive simulations and analysis of over 1,000,000 VAERS reports focusing on COVID-19 vaccine safety.

9:00-9:15 AM

Bayesian Generalized Linear Models for Compositional Data

Li Zhang, University of Alabama at Birmingham

We proposed Bayesian Generalized Linear Models for Compositional data (BGLMC) with a soft sum-to-zero restriction on coefficients through the prior distribution for the sum or mean of the coefficients. In the proposed structure, instead, to make sum/mean of coefficients exactly zero, we assume their mean follows a Normal distribution with mean zero and a small variance (e.g., 0.001). We combined the soft sum-to-zero restriction with various priors on the coefficients. The proposed method was implemented by R package Stan. The performance of proposed method was assessed by extensive simulation studies for both continuous and binary outcomes. The results of simulations show that our proposed method outperforms conventional methods without constraints in terms of variable selection and prediction. The proposed method was also applied to find microorganisms linked to inflammatory bowel disease (IBD) in two microbiome studies. To sum up, our proposed approach is capable of handling compositional data as well as high dimensionality for varied response distributions.

9:15-9:30 AM

Bayesian Predictive Modeling of Multi-Source Multi-Way Data

Jonathan Kim*, University of Minnesota

We develop a Bayesian approach to predict a continuous or binary outcome from data from multiple sources with a multi-way (i.e. multidimensional tensor) structure. As a motivating example we consider molecular data from multiple omics sources, each measured over multiple developmental time points, as predictors of early-life iron deficiency (ID) in a rhesus monkey model. We use a linear model with a low-rank structure on the coefficients to capture multi-way dependence and model the variance of the coefficients separately across each source to infer their relative contributions. Conjugate priors facilitate an efficient Gibbs sampling algorithm for posterior inference, assuming a continuous outcome with normal errors or a binary outcome with a probit link. Simulations demonstrate our model performs as expected in terms of misclassification rates and correlation of estimated coefficients with true coefficients, with large gains in performance by incorporating multi-way structure and modest gains when accounting for differing signal sizes across the different sources. Moreover, it provides robust classification of ID monkeys for our motivating application.

9:30-9:45 AM

Modeling Data Using Horseshoe Process Regression

Elizabeth Chase, University of Michigan

Biomedical data often exhibit jumps or abrupt changes. For example, women's basal body temperature may jump at ovulation and menstruation. These sudden changes make these data challenging to model: many methods will oversmooth the sharp changes or overfit in response to measurement error. We develop horseshoe process regression (HPR) to address this problem. We define a horseshoe process as a stochastic process in which each increment is horseshoe-distributed. We use the horseshoe process as a nonparametric Bayesian prior for modeling a potentially nonlinear association between an outcome and its continuous predictor, which we implement via Stan and in the R package HPR. We provide guidance and extensions to advance HPR's use in applied practice: we introduce a Bayesian imputation scheme to allow for interpolation at unobserved values of the predictor within the HPR; include additional covariates via a partial linear model framework; and allow for monotonicity constraints. We find that HPR performs well when fitting functions that have sharp changes. We apply HPR to model women's basal body temperatures over the course of the menstrual cycle.

9:45-10:00 AM

Reinforced Borrowing Framework: Leveraging Auxiliary Data for Individualized Inference

Ziyu Ji, Division of Biostatistics, University of Minnesota

During the past decade, it has been a common interest for researchers in many disciplines to leverage auxiliary data for enhancing individualized inference. Many statistical methods, such as multisource exchangeability models (MEM), are developed to borrow information from potentially heterogeneous supplementary sources to support the individual-level parameter inference. However, MEM and its alternatives only cover data from the parameter of interest and discard other information that may also contribute to indicating the exchangeability of supplementary sources. In this article, we propose a generalized Reinforced Borrowing Framework (RBF) using a distance-embedded prior within MEM which not only utilizes data about the parameter of interest but also uses different types of auxiliary information sources to improve the individualized parameter inference, with performance advantages and minimal additional computational burden. We demonstrate the application of RBF to COVID Travel Impact (CTI) Study to investigate the impact of the COVID-19 pandemic on people's individual activity and behaviors, where our approach achieves ~20% lower MSE compared with MEM.

10:00-10:15 AM

Bayesian Inference and Dynamic Prediction for Multivariate Longitudinal and Survival Data

Haotian Zou, University of North Carolina at Chapel Hill

Alzheimer's disease (AD) is a complex neurological disorder impairing multiple domains such as cognition and daily functions. To better understand the disease and its progression, many AD research studies collect multiple longitudinal outcomes that are strongly predictive of the onset of AD dementia. We propose a joint model based on a multivariate functional mixed model framework (referred to as MFMM-JM) that simultaneously models the multiple longitudinal outcomes and time to dementia onset. We develop six functional forms to fully investigate the complex association between longitudinal outcomes and dementia onset. Moreover, we use Bayesian methods for statistical inference and develop a dynamic prediction framework that provides accurate personalized predictions of disease progressions based on new subject-specific data. We apply the proposed MFMM-JM to two large ongoing AD studies: the Alzheimer's Disease Neuroimaging Initiative (ADNI) and National Alzheimer's Coordinating Center (NACC), and identified the functional forms with the best predictive performance. Our method is also validated by extensive simulation studies with five settings.

10. CONTRIBUTED PAPERS: HIGH DIMENSIONAL DATA ANALYSIS

Chair: Sarah Samorodnitsky, University of Minnesota

8:30-8:45 AM

Bayesian Indicator Variable Selection with High-Dimensional Multivariate Response for Detecting Noncoding RNA Regulators of Gene Expression

Hongjie Ke, University of Maryland, College Park

With the large amount of noncoding RNAs (ncRNAs) in human genome, their functions is, however, generally under-studied. Although the regulatory roles of ncRNAs in the process of gene expression has been widely pointed out, their regulatory mechanism has remained unclear. The high-dimensionality of both ncRNA data and gene expression data and their complex inter-feature correlation structure have posed a challenge in developing novel statistical and computational approaches. In addition, we wish to include the cancer stage information in the model which will enable us to identify the differential expressed genes and understand how the ncRNA regulatory mechanism changes over different stages of cancer. Here we propose a multivariate Bayesian indicator variable selection model that targets ncRNA regulation of gene expression problems. Our model includes the cancer stage information (main effect) and the interaction effect between ncRNA and cancer stage in order to study the differential regulated relationship. Simulation studies and application to multi-omics data demonstrate the validity and advantage of our method.

8:45-9:00 AM

Paired Analysis of Multi-Source TCGA Data Using JIVE

Michael O'Connell, Miami University

Many cancer genomics studies, including The Cancer Genome Atlas (TCGA), use healthy controls from the same subjects as the tumor samples. Although this is naturally paired data since the samples in each group (tumor and healthy) come from the same subjects, many analyses ignore the paired structure. This is partly influenced by incomplete data for the healthy controls, as the TCGA only has control data for a small subset of the total samples. This creates a dilemma of either ignoring the pairing in the data to use all of the tumor data or losing most of the tumor data to appropriately match the paired samples. In this project, we used the availability of multiple omics data sources to impute some of the missing controls, allowing a more comprehensive paired analysis of the data. Subsequently, we applied the Joint and Individual Variation Explained method to the differential omics matrices to explore the sources of variability in the paired data.

9:00-9:15 AM

Probabilistic Multilevel Canonical Correlation Analysis (CCA) for Integrative Analysis of Multi-Omics Data

Yuna Kim, Drexel University Dornsife School of Public Health

Multi-omics data have been used to characterize covariation in multiple biological profiles, allowing for a more comprehensive understanding of complex biological processes. Moreover, the reduction in costs of high-throughput technologies has further broadened the scope of multi-omics studies enabling the collection of repeated measurements or longitudinal data. While mixed effects models are widely used in single omics applications, their use in applications for integrative analyses of multi-omics data with a multilevel structure is less developed. Probabilistic canonical correlation analysis (pCCA) considers probability models for jointly studying the relations among two sets of data. We propose a probabilistic multilevel CCA that extends pCCA to multilevel data to help learn the underlying shared structures between two omics data sources simultaneously at both the within- and between-subject levels. We further extend the method for a variable selection and better interpretability by imposing sparsity on the feature loadings via the adaptive lasso. We examine our proposed method's operating characteristics and variable selection performance through simulation studies.

9:15-9:30 AM

High-Dimensional Mixed Graphical Models with Applications to Genomic Data Integration

Laurent Briollais, Lunenfeld-Tanenbaum Research Institute

Recent advances in biological research have seen the emergence of high-throughput technologies with numerous applications. In cancer research, the challenge is now to perform integrative analyses of high-dimensional multi-omic data with the goal to better understand genomic processes that correlate with cancer outcomes. We propose here a novel mixed graphical model approach to analyze multi-omic data of different types (continuous, discrete and count) and perform model selection by extending the Birth-Death MCMC (BDMCMC) algorithm. We compare the performance of our method to the LASSO and the standard BDMCMC methods using simulations and found that our method is superior in terms of both computational efficiency and the accuracy of the model selection results. Finally, an application to the TCGA breast cancer data shows that integrating genomic information at different levels (mutation and expression data) leads to better subtyping of breast cancers.

9:30-9:45 AM

Global Testing and Screening of Dynamic Effects

Ying Cui, Emory University

Identifying outcome relevant variables is generally of keen interest in biomedical studies. This task can be complicated by the consideration of dynamic (or varying) variable effects that often manifest meaningful scientific mechanisms. In this case, traditional approaches that perform tests oriented to outcome mean independence of the variables may be in jeopardy of depreciating some important variables. In this work, we propose a model-free testing and screening framework which adopts a global view pertaining to quantile independence to permit robust assessment of the effects which are possibly varying. Our testing procedure flexibly allows for evaluating multiple variables simultaneously and naturally evolves into unconditional and conditional screening procedures in ultra-high dimensional settings that possess the desirable sure screening property. We demonstrate good practical utility of our proposals via extensive simulation studies and a real application to a microarray data set.

9:45-10:00 AM

Joint Variable Selection of Functional Mixture Regression

Han Chen, Virginia Tech

We consider the variable selection problem in finite mixture of regression models (FMR) for high-dimensional data. FMR allows coefficients of covariates varying from one regression component to another. All regression components shared the same set of potentially important covariates, and thus common structure may exist in all components. To exploit the common structure among them, we proposed a joint variable selection method of functional mixture regression. Using an appropriate reparameterization, our method can extract a shared common structure across mixture models and also allow to identify individual structures for each regression model. Furthermore, we extended our method into mixture of functional linear models. Our joint variable selection method is shown to be consistent. A new EM algorithm is proposed for efficient numerical computation. Finally, we demonstrate the performance of our new method through simulation and real data analysis on market segmentation data and fMRI brain image data.

10:00-10:15 AM

Sufficient Dimension Reduction for Poisson Regression

Jianxuan Liu, Syracuse University

Poisson regression is popular and commonly employed to analyze frequency of occurrences in a fixed amount of time. In practice, data collected from many scientific disciplines tend to grow in both volume and complexity. One characteristic of such complexity is the inherent sparsity in high-dimensional covariates space. Sufficient dimension reduction (SDR) is known to be an effective cure for its advantage of making use of all available covariates. Existing SDR techniques for a continuous or binary response do not naturally extend to count response data. It is challenging to detect the dependency between the response variable and the covariates due to the curse of dimensionality. To bridge the gap between SDR and its applications in count response models, an efficient estimating procedure is developed to recover the central subspace through estimating a finite dimensional parameter in a semiparametric model. The proposed model is flexible and the resulting estimators achieve optimal semiparametric efficiency without imposing linearity or constant variance assumptions.

11. CONTRIBUTED PAPERS: CAUSAL INFERENCE IN OBSERVATIONAL AND LONGITUDINAL STUDIES

Chair: Fatema Shafie Khorassani, University of Michigan

8:30-8:45 AM

Doubly Robust Estimation of Causal Effects in Network-Based Observational Studies

Vanessa McNealis, Department of Epidemiology and Biostatistics, McGill University

Causal inference on populations embedded in social networks poses technical challenges, since the typical no interference assumption may no longer hold. For instance, in the context of infectious disease, the outcome of a study unit will likely be affected by the treatment of neighbours. While inverse probability weighted (IPW) estimators have been developed for this setting, they are often highly inefficient. In this work, we assume that the network is a union of connected subnetworks and propose doubly robust (DR) estimators combining models for treatment and outcome that are consistent and asymptotically normal if either model is correctly specified. We present empirical results that illustrate the DR property and the efficiency gain of DR over IPW estimators when both the outcome and treatment models are correctly specified. Simulations are conducted under different scenarios of (latent) treatment dependence. We apply these methods in an illustrative analysis using the Add Health data, re-examining a question first considered by previous authors to understand the impact of maternal college education on adolescent school performance, both directly and indirectly.

8:45-9:00 AM

Stable Estimation of Time-Varying Treatment Effects

Yige Li, Harvard University

In longitudinal studies, estimating time-varying treatment effects in the presence of time-varying confounders is widely discussed. Under certain assumptions, the potential outcomes under a known treatment path can be identified by g-formula or inverse probability of treatment weighting (IPTW). The weights in IPTW are typically estimated by explicitly modeling the probabilities of receiving treatment at each time point. However, in practice, the resulting weights may yield unstable treatment effect estimates and fail to balance covariates as they are supposed to. To address these problems, rather than explicitly modeling the probabilities of treatment, we directly find the weights of minimum variance that balance the covariates across all possible temporal treatment paths to their corresponding targets. In simulation studies, we show that this approach outperforms several existing methods by providing more precise time-varying treatment effect estimates. In addition, the proposed approach has comparative performance to g-formula estimation on efficiency yet is more robust to misspecification in the outcome model.

9:00-9:15 AM

An Efficient Propensity Score Method for Causal Analysis with Application to Case-Control Study in Breast Cancer Research

Azam Najafkouchak, Michigan State University

Propensity score has become a popular method to adjust for measured confounding factors in the absence of randomization. In real applications, a practice is to discretize these scores and use the stratification approach to estimate the causal parameter of interest. We show in this paper that such a practice is dangerous and may lead to bias and inefficient inferences, especially if there are continuous confounders. We introduce a novel and flexible stratification approach that uses all available information in the propensity score to improve the power to evaluate the average treatment effect. With continuous confounders, the approach does not rely on arbitrary discretizations. Instead, a scanning approach requiring continuous dichotomizations of the propensity score is proposed. Empirical processes resulting from these dichotomizations are then used to construct an integrated estimator of the causal effect, with limiting null distributions shown to be functionals of tight random processes. We illustrate our proposals using simulation studies and an application to a real data set in breast cancer research.

9:15-9:30 AM

Generalized Propensity Scores for Causal Analysis Involving Multiple Exposures

Kecheng Li, University of Waterloo

Propensity scores play a central role in modern causal inference, but related methods have primarily been directed at the setting of a univariate exposure variable. We describe the challenge of causal inference with multiple exposure variables, motivated by an epidemiological cohort study investigating the effects of prenatal alcohol exposure on child cognition. Accurate characterization of the dose-response surface is critical to aid in the diagnosis of children and the introduction of appropriate interventions and support. A generalized propensity score is developed based on a multivariate model for different dimensions of exposure (proportion of days spent drinking, average volume of alcohol per drinking day), which is used to define weights for inverse density weighted estimating functions and propensity-based covariates for regression adjustment. We investigate the importance of dependence modeling for multivariate exposures given confounders, and evaluate the implications of inadequately modeling the dependence structure. The methods are illustrated through application to data from our motivating study.

9:30-9:45 AM

Group Sequential Testing Under Instrumented Difference-in-Difference Approach

Samrat Roy, University of Pennsylvania

Unmeasured confounding is a major hindrance to reliable causal inferences based on observational studies. Instrumented difference-in differences (iDiD), a novel idea that connects the two well-known concepts of instrumental variable and standard DiD, ameliorates the above issue by explicitly leveraging exogenous randomness in an exposure trend. The setup and assumptions of standard DiD originate from the applications in social sciences, and it is difficult to find examples in biomedical sciences where the assumptions of DiD are met. As a remedy, the newly developed idea of iDiD exploits a haphazard encouragement referred to as IV for DiD, which is targeted at a subpopulation towards faster uptake of the exposure. In this article, we utilize the above idea of iDiD, and propose a novel group sequential testing method that provides valid inference even in the presence of unmeasured confounders. The performance of our proposed approach is evaluated on both synthetic data and Clinformatics Data Mart Database to examine the association between rofecoxib and acute myocardial infarction (AMI).

9:45-10:00 AM

Flexible Template Matching for Observational Study Design

Bo Lu, The Ohio State University

Matching is a popular design for inferring causal effect with observational data. The application of matched design for real world data may be limited by: 1) the causal estimand of interest; 2) the sample size of different treatment arms. We propose a flexible design of matching, based on the idea of template matching, to overcome these challenges. It first identifies the template group which is representative of the target population, then match subjects from the original data to this template group and make inference. It can unbiasedly estimate the ATE using matched pairs or estimate the ATT when the treatment group has a bigger sample size. One major advantage of matched design is that it allows both randomization-based or model-base inference, with the former being more robust. For the commonly used binary outcome in medical research, we adopt a randomization inference framework of attributable effects in matched data, which allows heterogeneous effects and can incorporate sensitivity analysis for unmeasured confounding. We apply our design and analytical strategy to a trauma care evaluation study.

10:00-10:15 AM

Estimation of Average Treatment Effect for Survival Outcomes with Continuous Treatment in Observational Studies

Triparna Poddar, University of Louisville

Recent literature on causal effect for survival analyses mainly focus on multiple treatment settings, studies with continuous treatment setting are seldom explored. Here, we estimate the average treatment effect (ATE) of continuous treatment on time to event outcomes by adjusting multiple confounding factors and considering censored observations. To adjust confounding factors, various propensity score methods such as multinomial regression and covariate balance propensity score models are used to estimate the ATE via the inverse probability of treatment weighting (IPTW) method. For continuous treatments, the IPTW is generated from the covariate balancing generalized propensity score. To remedy the possible bias for time-to-event data with censored observations, we incorporate the censoring weights. We propose using both IPTW and censoring weights (say, double weighting) to estimate ATE using the marginal structural accelerated failure time (AFT) model. Extensive simulation studies demonstrated our proposed method performed well. We applied our proposed method to study the impact of lack of health insurance on the survival of patients diagnosed with alcoholic cirrhosis.

12. CONTRIBUTED PAPERS: METHODS FOR RANDOMIZED CLINICAL TRIALS, SURROGATE MARKERS

Chair: Ann Marie Weideman, University of Michigan

8:30-8:45 AM

Randomized Controlled Dose-Escalation Design to Evaluate the Safety of a Novel Pharmacological Cardiopulmonary Resuscitation Strategy

Sydney Benson, University of Minnesota

The motivating trial evaluates a novel pharmacological cardiopulmonary resuscitation technique to improve outcomes for out-of-hospital cardiac arrest patients through increased end organ perfusion. The proposed phase I trial design expands upon traditional dose-finding designs to include a randomized control arm to assess safety via serum lactate on hospital admission. We propose and compare six Bayesian models for guiding dose escalation. Each model makes different assumptions about the change in serum lactate across control cohorts. Model selection aims to minimize the expected number of incorrect dose-escalation decisions while sample size selection targets an expected number of incorrect decisions. Randomization is 1:1 for the initial cohort and for subsequent cohorts is chosen to maximize the lower confidence bound for the posterior precision of expected log serum lactate among controls in later cohorts versus the initial cohort. We find that the spike and slab model performs best. We also determine that cohorts 2 and 3 should use a 2:1 randomization ratio. On average, with this allocation, 70 individuals will ensure only 1 incorrect dose-escalation decision of 6.

8:45-9:00 AM

Use of Natural History Data for Drug Evaluation in Duchenne Muscular Dystrophy: A Bayesian Small Sample, Sequential, Multiple Assignment Randomized Trial Design

Sidi Wang, University of Michigan

In Duchenne muscular dystrophy (DMD) and other rare diseases, recruiting patients into clinical trials is challenging. Additionally, assigning patients to long-term, multi-year placebo arms raises ethical issues. Ideally, researchers could use natural history data that complements or enriches the placebo group. We propose a small sample, sequential, multiple assignment, randomized trial (snSMART) design that integrates external control data. The proposed approach is a multi-stage design evaluating multiple doses of a promising drug vs. placebo. To efficiently estimate treatment effects in a snSMART design, we present a robust MAC-snSMART, a robust exchangeable hierarchical model. We reanalyze a DMD trial using the proposed method and external control data from the Duchenne Natural History Study. Our method's estimators show improved efficiency compared to the original trial. Also, the robust MAC-snSMART provides more accurate estimators than the traditional analytic method when its assumptions (practical in most snSMART regimes) are not violated. Thus, the proposed design and method are a promising alternative tool for drug development in DMD and other rare diseases.

9:00-9:15 AM

Randomized Phase II Selection Design with Order Constrained Strata

Yi Chen*, University of Wisconsin Madison

The exploratory nature of phase II trials makes it quite common to include heterogeneous patient subgroups with different prognoses in the same trial. Incorporating such patient heterogeneity or stratification into statistical calculation for sample size can improve efficiency and reduce sample sizes in single-arm phase II trials. However, such consideration is lacking in randomized phase II trials. In this paper, we propose methods that can utilize some natural order constraints that may exist in stratified population to gain statistical efficiency for randomized phase II designs. For thoroughness and simplicity, we focus on the randomized phase II selection designs in this paper, although our method can be easily generalized to the randomized phase II screening designs. We consider both binary and time-to-event outcomes in our development. Compared with methods that do not use order contraints, our method is shown to improve the probabilities of correct selection in our simulated and real examples.

9:15-9:30 AM

A Semiparametric Cox-Aalen Transformation Model with Censored Data

Xi Ning*, University of North Carolina at Charlotte

We propose a broad class of so-called Cox-Aalen transformation models that incorporate both multiplicative and additive covariate effects on the baseline hazard function within a transformation. The proposed models provide a highly flexible and versatile class of semiparametric models that include the transformation models and the Cox-Aalen model as special cases. We propose an estimating equation approach and devise an Expectation-Solving (ES) algorithm that involves fast and robust calculations. The resulting estimator is shown to be consistent and asymptotically normal via modern empirical process techniques. The ES algorithm yields a computationally simple method for estimating the variance of both parametric and nonparametric estimators. Finally, we demonstrate the performance of our procedures through extensive simulation studies and applications in two randomized, placebo-controlled HIV prevention efficacy trials. The data example shows the utility of the proposed Cox-Aalen transformation models in enhancing statistical power for discovering covariate effects.

9:30-9:45 AM

Simultaneous Hypothesis Testing for Multiple Competing Risks in Comparative Clinical Trials

Jiyang Wen*, Johns Hopkins Bloomberg School of Public Health

Competing risks data are commonly encountered in randomized clinical trials or observational studies. Ignoring competing risks in survival analysis leads to biased risk estimates and improper conclusions. Often, one of the competing events is of primary interest and the rest competing events are handled as nuisances. These approaches can be inadequate when multiple competing events have important clinical interpretations and thus of equal interest. For example, in COVID-19 in-patient treatment trials, the outcomes of COVID-19 related hospitalization are either death or discharge from hospital, which have completely different clinical implications and are of equal interest, especially during the pandemic. In this paper we develop nonparamteric estimation and simultaneous inferential methods for multiple cumulative incidence functions (CIFs) and corresponding restricted mean times. Based on Monte Carlo simulations and a data analysis of COVID-19 in-patient treatment clinical trial, we demonstrate that the proposed method provides global insights of the treatment effects across multiple endpoints.

9:45-10:00 AM

Accounting for Inconsistent Use of Covariate Adjustment in Group Sequential Trials

Marlena Bannick, University of Washington

Group sequential designs are common in clinical trials, as they allow for interim efficacy and futility monitoring. Adjustment for baseline covariates can increase power and precision of estimated effects in clinical trials. Statistical methods for both group sequential trials and covariate adjustment are well-developed, including when covariate adjustment is used in a group sequential trial. What is less studied is how to perform interim monitoring, estimation, and inference in group sequential trials where covariate adjustment is applied inconsistently throughout the trial (e.g., if there is a delay in covariate data being made available). We show how to bridge this gap by studying the asymptotic behavior of test statistics obtained from a group sequential trial using ANOVA versus ANCOVA. We focus on two-arm trials with simple, balanced randomization and continuous outcomes. We study the performance of our boundary, estimation, and inference adjustments in simulation studies. We end with recommendations about the application of covariate adjustment in group sequential designs.

10:00-10:15 AM

Surrogacy Validation for Time-to-Event Outcomes with Illness-Death Frailty Models

Emily Roberts, University of Iowa

It is common in clinical trials to evaluate a treatment effect on an intermediate endpoint when the true outcome of interest would be difficult or costly to measure. We consider how to validate intermediate endpoints in a causally-valid way when the trial outcomes are time-to-event. Using counterfactual outcomes, those that would be observed if the counterfactual treatment had been given, the causal association paradigm assesses the relationship of the treatment effect on the surrogate with the treatment effect on the true endpoint. In particular, we propose illness death models to accommodate the censored and semi-competing risk structure of survival data. The proposed causal version of these models involves estimable and counterfactual frailty terms. Via these multi-state models, we characterize what a valid surrogate would look like using a causal effect predictiveness plot. We evaluate the estimation properties of a Bayesian method using Markov Chain Monte Carlo and assess the sensitivity of our model assumptions. Our motivating data source is a localized prostate cancer clinical trial where the two survival endpoints are time to distant metastasis and time to death.

Monday, March 20, 2023 | 10:30-12:15 PM

13. NEW FUNCTIONAL DATA METHODS FOR SURVIVAL DATA

Organizer: Luo Xiao, North Carolina State University
Chair: Salil Koner, Duke University

10:30-10:55 AM

The Functional Cox Model

Ciprian Crainiceanu, Johns Hopkins University

We propose the Functional Cox Model and its generalizations to flexibly quantify the association between functional covariates and time to event data. The model extends the linear proportional hazards model by allowing the predictors to be high dimensional functions and for the model structure to allow both linear and nonparametric smooth effects. We also discuss model identifiability and practical ways of inducing it into software. Methods are applied to the National Health and Nutrition Examination Survey (NHANES) 2003-2006 accelerometry data and quantify new and interpretable circadian patterns of physical activity that are associated with all-cause mortality. We also introduce a simple and novel simulation framework for generating survival data with functional predictors which resemble the observed data. Software will be demonstrated on simulated and the NHANES data.

10:55-11:20 AM

Functional Data Analysis for Longitudinal Data with Informative Observation Times

Luo Xiao, North Carolina State University

In functional data analysis for longitudinal data, the observation process is typically assumed to be noninformative, which is often violated in real applications.

Thus, methods that fail to account for the dependence between observation times and longitudinal outcomes may result in biased estimation. For longitudinal data with informative observation times, we find that under a general class of shared random effect models, a commonly used functional data method may lead to inconsistent model estimation while another functional data method results in consistent and even rate-optimal estimation. Indeed, we show that the mean function can be estimated appropriately via penalized splines and that the covariance function can be estimated appropriately via penalized tensor-product splines, both with specific choices of parameters. For the proposed method, theoretical results are provided, and simulation studies and a real data analysis are conducted to demonstrate its performance.

11:20-11:45 AM

Dynamic Prediction with Multivariate Longitudinal Outcomes and Longitudinal Magnetic Resonance Imaging Data

Sheng Luo, Duke University

Alzheimer's Disease (AD) is a common neurodegenerative disorder impairing multiple domains. Recent AD studies, the Alzheimer's Disease Neuroimaging Initiative (ADNI) study included, collect multimodal data to better understand AD severity and progression. It is essential to develop an AD predictive model that leverages multimodal data and provides accurate personalized predictions of dementia occurrence. In this article, we propose a multivariate functional mixed model with longitudinal magnetic resonance imaging data (MFMM-LMRI) that jointly models neurological scores, longitudinal voxel-wise MRI data, and the survival outcome as dementia onset. We investigate two functional forms linking the longitudinal and survival process and model longitudinal MRI data using joint and individual variation explained (JIVE) approach. We adopt Markov Chain Monte Carlo (MCMC) method to obtain posterior samples. We establish a dynamic prediction framework that predicts longitudinal trajectories, MRI data, and the probability of dementia occurrence. This methodology development is motivated by and applied to the ADNI study.

11:45-12:10 PM

Supervised FPCA and Long-Term Risk Prediction in Breast Cancer

Shu Jiang, Washington University School of Medicine

Screening mammography aims to identify breast cancer early and secondarily measures breast density to classify women at higher or lower than average risk for future breast cancer in the general population. Our primary goal in this study is to extract mammogram based features that augment the well-established breast cancer risk factors to improve prediction accuracy. In this talk, I will present a novel supervised functional principal component analysis to extract image-based features that are ordered by association with the failure times.

14. CAUSAL INFERENCE METHODS FOR ASSESSING THE EFFECT OF ENVIRONMENTAL AND NUTRITIONAL EXPOSURES ON PUBLIC HEALTH

Organizer/Chair: Nandita Mitra, University of Pennsylvania

10:30-10:55 AM

Estimating and Forecasting the Causal Effects of Extreme Weather Events on Health

Rachel Nethery, Harvard T.H. Chan School of Public Health

To minimize the health threats presented by extreme weather events, we must generate high-precision insights and tools to inform strategic preparedness efforts. Currently, our limited understanding of the epidemiology of these events inhibits progress in reducing health risks. We propose an integrated causal and predictive statistical modeling approach that, when applied to today's wealth of historic weather and health data, enables standardized, high-resolution quantification of the health impacts of historic extreme weather episodes and characterizes how features of the events and the impacted communities explain variation in health risks. This method enables high-resolution prediction of future extreme weather-related health impacts, which can inform strategic preparedness and aid in identifying high-risk communities in advance of future events. We apply our method to a rich data platform containing detailed historic tropical cyclone exposure information for the US and Medicare claims data to investigate health effects of past tropical cyclones and identify features predictive of tropical cyclone-related health risks.

10:55-11:20 AM

A Difference-in-Differences Framework to Estimate Causal Effects for Policy Interventions in the Presence of Heterogeneous Interference

Gary Hettinger, University of Pennsylvania

Public policy interventions are often evaluated with the difference-in-differences (DiD) approach, which does not directly account for a policy affecting nearby regions, especially when these effects vary spatially. For example, an excise tax on sweetened beverages in Philadelphia (PHL) was associated with substantial decreases in volume sales of taxed beverages in PHL as well as increases in beverage sales of nontaxed bordering counties. The latter association may be explained by cross-border shopping behaviors of PHL residents, which may vary with border proximity, transportation access, and demographics. Because such effects can offset the total effect of such interventions, particularly for specific sub-populations, understanding effect dynamics is essential to holistically evaluate public policies. Further, such insights may help predict policy effects under diverse implementation strategies. To address these concerns, we extend DiD methodology to robustly identify the causal effects of policy interventions under potentially heterogeneous interference exposure. Here, we present initial work demonstrating our framework with an evaluation of the PHL Beverage Tax policy.

11:20-11:45 AM

Longitudinal G-Computation for Policy Evaluation with Clustered Data

Nicholas Illenberger, NYU Langone Health

To evaluate the impact of time-varying policy interventions using observational panel data, researchers must carefully consider the effects of time-varying confounding. For example, diabetes rates within a state may influence policies affecting access to the Supplemental Nutrition Assistance Program (SNAP) and, consequently, these interventions may influence future diabetes rates and further policy interventions. Longitudinal g-computation provides a framework for estimating the causal effect of an exposure from observational data when time-varying confounding may be present. However, methods for performing g-computation when data exhibit natural clustering (e.g. county level diabetes rates and state level SNAP policies) have not been explored. We propose novel methodology for performing g-computation with clustered data. This approach accounts for within-cluster correlation and time-varying confounding while accommodating flexible modelling approaches. Through simulations, we assess the operating characteristics of our estimator under several common clustered data scenarios. This approach is used to assess the effect of SNAP policies on county-level diabetes rates.

11:45-12:10 PM

Bayesian Kernel Machine Regression for Environmental Mixtures

Linda Valeri, Columbia University

The impact of toxic chemical mixtures on the development of chronic disease is a critical public health concern. Knowledge of time windows of susceptibility and mechanisms explaining the harmful effects can help inform treatment and prevention strategies. Several factors challenge internal and external validity of health effect estimation when multiple exposures are time-varying: multi-collinearity, time-varying confounding, complex exposure response relationships, and exposure-covariates interactions. We develop a Bayesian Kernel Machine Regression approach for flexible estimation of g-formula (gBKMR) and mediation (BKMR-CMA). Our approach ranks importance of mixture components and exposure windows, accounts for non-linear and non-additive effects, adjusting for time-varying confounding. Simulation studies demonstrate the performance of our approaches in estimating direct, indirect, and time-varying effects under realistic exposure-response scenarios.We applied this methodology to quantify the contribution of birth length as a mediator between in utero co-exposure of arsenic, manganese and lead, and children neurodevelopment, in a prospective cohort in rural Bangladesh.

15. RECENT STATISTICAL ADVANCES IN MICROBIOME DATA ANALYSIS

Organizer/Chair: Tianying Wang, Tsinghua University

10:30-10:55 AM

MarZIC: A Marginal Mediation Model for Zero-Inflated Compositional Mediators with Applications to Microbiome Data

Meilin Jiang, University of Florida

Standard mediation analysis methods are not adequate to analyze the microbiome as a mediator due to the excessive number of zero-valued sequencing reads in the data that is compounded by its compositional structure. The two main challenges raised by the zero-inflated data structure are: (a) disentangling the mediation effect induced by the point mass at zero; and (b) identifying the observed zero-valued data points that are not zero (i.e., false zeros). We develop a novel marginal mediation analysis method under the potential-outcomes framework to fill this gap and show the marginal model can also account for the compositional structure. The mediation effect can be decomposed into two components that are inherent to the two-part nature of zero-inflated distributions. With probabilistic models to account for observing zeros, we also address the challenge with false zeros. A comprehensive simulation study and the application in a real microbiome study showcase our approach in comparison with existing approaches.

10:55-11:20 AM

ODE-Based Deep Learning Interpolation for Irregularly-Sampled Longitudinal Microbiome Data

Di Wu, University of North Carolina at Chapel Hill

Microbiome has been studied in many biomedical areas. Longitudinal microbiome data are increasingly collected recently. However, most samples in a large population of longitudinal studies are irregularly-sampled that are not collected at the exactly same time unit. Microbiome data are typically zero-inflated and over-dispersed. Existing methods hardly capture these features and may weaken downstream analysis. We proposed a deep-learning-based interpolation model called Bidirectional GRU-ODE-Bayes (BGOB), specifically for the large longitudinal microbiome datasets. BGOB can be applied to both univariate and multivariate data, as well as compositional and count data. Data preprocessed with BGOB is flexible in regression, classification and clustering et al. Simulations demonstrate that our model outperforms other models in regression, classification, and clustering. Application in the human oral microbiome generates the clusters of taxa and clusters of samples based on the time trajectory.

11:20-11:45 AM

Testing Microbiome Associations with Survival Times at Both the Community and Individual Taxon Levels

Yijuan Hu, Emory University

Finding microbiome associations with possibly censored survival times is an important problem, especially as specific taxa could serve as biomarkers for disease prognosis or as targets for therapeutic interventions. The existing methods MiRKAT-S and OMiSA are restricted to testing the associations at the community level and do not provide results at the individual taxon level. An ad hoc approach testing each taxon using the Cox model may not perform well with sparse taxa count data and small sample sizes. We have previously developed the linear decomposition model (LDM) for testing continuous or discrete outcomes that unifies community-level and taxon-level tests into one framework. Here we extend the LDM to test survival outcomes, by using the Martingale (or deviance) residuals obtained from the Cox model as continuous covariates in the LDM. Using simulated data, we showed that the LDM-based tests preserved FDR for testing individual taxa and had good sensitivity. An analysis of data on the association of the gut microbiome and the time to acute graft-versus-host disease revealed several dozen associated taxa that would not have been achievable by any community-level test.

11:45-12:10 PM

A Semiparametric Quantile Single-Index Model for Zero-Inflated and Over-dispersed Outcomes

Tianying Wang, Center for Statistical Science, Tsinghua University

We consider the complex data modeling problem motivated by the zero-inflated and over-dispersed microbiome read count data. Several parametric approaches have been proposed to address issues of zero inflation and overdispersion, such as zero-inflated Poisson regression and zero-inflated Negative Binomial regression. However, parametric assumptions could be easily violated in real-world applications. To relax the parametric assumptions and provide a robust modeling framework, we propose a semiparametric single-index quantile regression framework, which is flexible to include a wide range of possible association functions and adaptable to the various zero proportions across subjects. We establish the asymptotic normality of the index coefficients estimator and the asymptotic convergence rate of the nonparametric quantile regression curve estimation. Through Monte Carlo simulation studies and the application in a microbiome study, we demonstrate the superior performance of the proposed method.

16. STATISTICAL METHODS FOR DOSE OPTIMIZATION IN ONCOLOGY TRIALS

Organizer: Yuan Ji, The University of Chicago
Chair: Kentaro Takeda, Astellas Pharma

10:30-10:55 AM

A Semi-Mechanistic Dose-Finding Design in Oncology Using Pharmacokinetic/Pharmacodynamic Modeling

Yisheng Li, The University of Texas MD Anderson Cancer Center

The commonly used phase I dose-finding designs in oncology are either algorithmic or empirical model-based. We propose a new framework for modeling the dose-toxicity relationship, by incorporating the pharmacokinetic (PK) data and hypothesized mechanisms of the drug effects, via dynamic PK and latent PD modeling. Our simulation studies show, with moderate departure from the hypothesized mechanisms of the drug action, that the performance of the proposed design on average improves upon those of the common designs, including the continual reassessment method, Bayesian optimal interval design, modified toxicity probability interval method, and a design called PKLOGIT that models the effect of the area under the concentration-time curve on toxicity. In case of considerable departure from the underlying drug effect mechanism, the performance of the design is shown to be comparable with those of the other designs. We illustrate the proposed design by applying it to a phase I trial of a Î³-secretase inhibitor in metastatic or locally advanced solid tumors. We also provide R code to implement the proposed design.

10:55-11:20 AM

Dose Optimization and Selection for Early Oncology Clinical Development - A Randomized Phase I-II Distributed Dose-Finding Design Scheme

Haitao Pan, St. Jude Children's Research Hospital

The FDA Oncology Center of Excellence (OCE) Project Optimus is an initiative to reform the dose optimization and dose selection paradigm in oncology drug development. One of the specific goals of Project Optimus is to develop strategies for dose finding and dose optimization across oncology that emphasizes the selection of a dose or doses that maximize not only the efficacy of a drug but the safety and tolerability as well, including randomized evaluations of a range of doses in trials. To this end, we introduce a phase I/II seamless dose escalation/expansion with an adaptive randomization scheme (SEARS), which provides a fit-for-purpose structure. SEARS is a seamless design that combines phase I dose escalation based on toxicity with phase II dose expansion and dose comparison based on efficacy. SEARS allows extension from phase I to phase II in a single protocol with no gap in between and employs a dynamic and parallel procedure involving simultaneous dose escalation, dose selection, and adaptive randomization. An R package SEARS demonstrates how to implement the proposed design in practice.

11:20-11:45 AM

Probability-of-Decision Designs to Accelerate Dose-Finding Trials

Tianjian Zhou, Colorado State University

Cohort-based enrollment can slow down phase I dose-finding trials since the outcomes of the previous cohort must be fully evaluated before the next cohort can be enrolled. This results in frequent suspension of patient enrollment. We propose a class of probability-of-decision (POD) designs to accelerate dose-finding trials, which enable dose assignments in real-time in the presence of pending toxicity outcomes. With uncertain outcomes, the dose assignment decisions are treated as random variables, and we calculate the posterior distribution of the decisions. The posterior distribution reflects the variability in the pending outcomes and allows a direct and intuitive evaluation of the confidence of all possible decisions. Optimal decisions are calculated based on the 0-1 loss, and extra safety rules are constructed to enforce sufficient protection from exposing patients to risky doses. A new and useful feature of POD designs is that they allow investigators and regulators to balance the trade-off between enrollment speed and making risky decisions by tuning a pair of intuitive design parameters. The performances of POD designs are evaluated through numerical studies.

11:45-12:10 PM

DISCUSSANT

Yuan Ji, University of Chicago

17. RANDOMIZATION IN CLINICAL TRIAL DESIGN AND INFERENCE: REDUCING CHANCE IMBALANCE AND ADDRESSING TRIAL DISRUPTIONS

Organizer: Jonathan Chipman, University of Utah
Intermountain Healthcare
Chair: Frank Bretz, Novartis

10:30-10:55 AM

Selecting an Appropriate Randomization Procedure for a Randomized Controlled Trial: A Statistician's Perspective

Oleksander Sverdlov, Novartis

The randomized controlled trial (RCT) is an established gold standard design for evidence-based medicine. Various kinds of randomization procedures are available, and choosing an “optimal” procedure for a given trial is not straightforward. In this talk we will present a systematic roadmap for the choice and application of a restricted randomization procedure in a randomized, comparative, parallel group clinical trial with equal (1:1) allocation. Important statistical considerations will be highlighted, including a tradeoff between treatment balance and allocation randomness, type I error, and power of a statistical test. Some real-life clinical trial examples will be presented to illustrate the thinking process for selecting a randomization procedure for implementation in practice. We will argue that randomization-based tests are robust and valid alternatives to likelihood-based tests and should be considered more frequently by clinical investigators.

10:55-11:20 AM

Experimenting with Finite to Infinite Population Sizes

Jonathan Chipman, University of Utah Intermountain Healthcare

ANOVA is commonly used to compare group means in randomized trials and assumes random sampling from an infinite population. However, it has been argued that a trial's participants reflect a finite population. A finite population Central Limit Theorem provides a degree of reassurance to assume normality when carrying out complete randomization with fixed equal allocation (i.e. random allocation rule, RAR). In practice, many trials further reduce the risk of chronological bias by using a maximum tolerable imbalance procedure (MTI) or permuted block design (PBD). Through extensive simulation, we investigate the impact upon Type I error when using RAR, MTI, or PBD for a full or partial sample of a finite population. Trials that implement MTI and PBD require a larger sample size than RAR to confidently control Type I error. Adjusting for design parameters, such as block assignment in PBD, further helps control Type I error. Randomization-based inference ensures exact inference under all settings, though causal testing is limited to the observed participants.

11:20-11:45 AM

Randomization Tests in Clinical Trials with Multiple Imputation for Handling Missing Data

Anastasia Ivanova, Department of Biostatistics, UNC at Chapel Hill

Randomization-based inference is a useful alternative to traditional population model-based methods. In trials with missing data, multiple imputation is often used. We describe how to construct a randomization test in clinical trials where multiple imputation is used for handling missing data. We illustrate the proposed methodology using Fisher’s combining function applied to individual scores in two post-traumatic stress disorder trials.

11:45-12:10 PM

Randomization Tests to Address Disruptions in Clinical Trials

Sergey Tarima, Medical College of Wisconsin

The COVID-19 pandemic had numerous consequences for ongoing clinical trials. People around the globe restricted their daily activities to minimize contagion, which led to missed visits and cancelling or postponing of elective medical treatments. For some clinical indications, COVID-19 may lead to a change in the patient population or treatment effect heterogeneity. We will measure the effect of the disruption on randomization tests and derive a methodological framework for randomization tests that allows for the assessment of clinical trial disruptions. We show that randomization tests are robust against clinical trial disruptions in certain scenarios, namely if the disruption can be considered an ancillary statistic to the treatment effect. As a consequence, randomization tests maintain type I error probability and power at their nominal levels

18. MACHINE LEARNING WITH APPLICATIONS TO EMERGING TOPICS IN BIOMEDICAL SCIENCE

Organizer: Cai Li, St. Jude Children's Research Hospital
Chair: Zilin Li, Indiana University School of Medicine

10:30-10:55 AM

Machine Learning with Time-to-Event Data: Estimating Conditional Survival Curves

Noah Simon, University of Washington, Biostatistics

It is of increasing interest to build predictive models for time-to-event outcome data, using potentially complex features (eg. images, ECGs, pathologic slides). There are two related tasks here: Risk stratification and conditional survival estimation. Under certain assumptions, eg. proportional hazards, these two tasks coincide. In general however, they do not: Conditional survival estimation is more ambitious. In this work we discuss a framework that relates conditional survival estimation to probabilistic classification: This allows us to leverage ML tools for classification in the survival domain. This framework can also accommodate censoring and truncation.

10:55-11:20 AM

Over-Sampling Matching Learning: Deriving Individualized Treatment Rules using Observational Data with Unbalanced Treatment Groups

Yiwang Zhou, St. Jude Children's Research Hospital

Precision medicine is gaining increasing attention in biomedical sciences. Spinal muscular atrophy (SMA) is a neurodegenerative disorder presenting usually in infancy and childhood. The motor skills of SMA patients are limited, which may be treated by Spinraza that specifically targets the generation of the survival motor neuron protein. This project aims to establish an individualized treatment rule (ITR) that can guide SMA patients on taking Spinraza to maximize their motor function. We propose a new statistical learning method, termed over-sampling matching learning (OM-learning), to address two major challenges in the ITR derivation using observational data, including the unbalanced covariate distributions and the unbalanced sample sizes for treatment groups. OM-learning first over-samples the minority class with the generation of synthetic examples using the kernel-based synthetic minority over-sampling technique. Then an ITR is derived based on matching learning. Applying OM-learning to the SMA study, we identified several useful biomarkers to form an ITR that would help maintain a higher motor function should it be implemented in the whole study population.

11:20-11:45 AM

SEAGLE: A Scalable Exact Algorithm for Large-Scale Set-Based GxE Tests in Biobank Data

Jocelyn Chi, UCLA

The explosion of biobank data offers immediate opportunities for gene-environment (GxE) interaction studies of complex diseases because of the large sample sizes and the rich collection in genetic and non-genetic information. However, the extremely large sample size also introduces new computational challenges in GxE assessment, especially for set-based GxE variance component (VC) tests, which are a widely used strategy to boost overall GxE signals and to evaluate the joint GxE effect of multiple variants from a biologically meaningful unit (e.g., gene). In this work, we focus on continuous traits and present SEAGLE, a Scalable Exact AlGorithm for Large-scale set-based GxE tests, to permit GxE VC tests for biobank-scale data. SEAGLE employs modern matrix computations to achieve the same "exact" results as the original GxE VC tests without imposing additional assumptions or relying on approximations. SEAGLE can easily accommodate sample sizes in the order of 105, is implementable on standard laptops, and does not require specialized computing equipment. We demonstrate SEAGLE's performance through extensive simulations. We illustrate its utility by conducting genome-wide gene-based GxE analysis on the Taiwan Biobank data to explore the interaction of gene and physical activity status on body mass index.

11:45-12:10 PM

Distribution-Invariant Differential Privacy

Xuan Bi, University of Minnesota

Differential privacy is becoming one gold standard for protecting the privacy of publicly shared data. It has been widely used in biomedical sciences, data science, public health, information technology, and the U.S. decennial census. Nevertheless, to guarantee differential privacy, existing methods may unavoidably alter the conclusion of the original data analysis, as privatization often changes the sample distribution. This phenomenon is known as the trade-off between privacy protection and statistical accuracy. In this work, we mitigate this trade-off by developing a distribution-invariant privatization (DIP) method to reconcile both high statistical accuracy and strict differential privacy. As a result, any downstream statistical or machine learning task yields essentially the same conclusion as if one used the original data. Numerically, under the same strictness of privacy protection, DIP achieves superior statistical accuracy across a wide range of simulation studies and real-world benchmarks.

19. CONTRIBUTED PAPERS: CLINICAL TRIAL METHODS

Chair: Holly Hartman, Case Western Reserve University

10:30-10:45 AM

Semi-Parametric Sensitivity Analysis for Trials with Irregular and Informative Assessment Times

Bonnie Smith, Johns Hopkins Bloomberg School of Public Health

Many trials are designed to collect outcomes at or around pre-specified times after randomization. In practice, there can be substantial variability in the times at which participants are actually assessed. When comparing outcomes at the (random) times of assessment, differences in treatment arms can be driven by the timing of assessments, rather than by an effect of treatment on underlying outcome trajectories. To avoid this problem, one can focus on the effect of treatment at each of the (fixed) targeted assessment times. However, untestable assumptions are needed, so it is important to assess how inferences would change under departures from these assumptions. We develop a sensitivity analysis methodology for this setting, along with a semi-parametric, influence function-based estimation approach. Our method allows for flexible modeling of the intensity function and outcome regression, while yielding an estimator for the target parameter that converges at root-n rates. We apply our method to a study of participants with uncontrolled asthma.

10:45-11:00 AM

Group Sequential Two-Stage Preference Designs

Ruyi Liu, Department of Biostatistics, Yale Center for Analytical Sciences, Yale School of Public Health

The two-stage preference design (TSPD) studies treatment efficacy while incorporating patient preference to treatment, providing unbiased estimates of selection and preference effects. One potential barrier to adopting TSPD in practice is the relatively large sample size required to estimate selection and preference effects with sufficient power. To address this limitation, we propose a novel group sequential two-stage preference design (GS-TSPD), which combines TSPD with sequential monitoring. In GS-TSPD, the pre-planned sequential monitoring allows investigators to conduct repeated hypothesis tests on accumulated data before complete enrollment. Without inflating the type I error, investigators terminate the study when sufficient evidence of the test effect is identified during an interim analysis, therefore reducing the design resource in expectation. We establish the independent increments assumption for selection and preference to approximate sequential density functions and apply sequential stopping boundaries. Simulations are conducted to establish the efficiency and operating characteristics of our proposed GS-TSPD, and then apply it to a hepatitis C virus study.

11:00-11:15 AM

An Optimal Two-Period Multiarm Platform Design with New Experimental Arms Added During the Trial

Xiaomeng Yuan, St. Jude Children's Research Hospital

Platform trials are multiarm clinical studies that allow the addition of new experimental arms after the activation of the trial. Statistical issues, with respect to ?adding new arms?, however, have not been thoroughly discussed. This work was motivated by a ?two-period? pediatric osteosarcoma study, which will start with two experimental arms and later add two more pre-planned experimental arms. In this study, we provide a principled approach, including, how to modify the critical boundaries to control the family-wise error rate when new arms are added, how to re-estimate the sample size and provide the optimal control-to-experimental arms allocation ratio, in terms of minimizing the total sample size to achieve a desirable power. The influence of the timing when new arms are added on the design?s operating characteristics is also examined, which provides a practical guide for deciding the timing of adding new arms. Other various numerical evaluations have also been conducted. We have developed a R package, PlatformDesign, for practitioners to easily use this platform trial approach.

11:15-11:30 AM

Estimating the Distribution of Growth Modulation Index in Phase II Oncology Trials

Li Chen, University of Kentucky

With the rapid development of new anti-cancer agents which are cytostatic, new endpoints are needed to better measure treatment efficacy in phase II trials. For this purpose, Von Hoff (1998) proposed the growth modulation index (GMI), i.e. the ratio between times to progression or progression-free survival times in two successive treatment lines. An essential task in studies using GMI as an endpoint is to estimate the distribution of GMI. Traditional methods for survival data have been used for estimating the GMI distribution because censoring is common for GMI data. However, we point out that the independent censoring assumption required by traditional survival methods is always violated for GMI, which may lead to severely biased results. In this paper, we construct both nonparametric and parametric estimators for the distribution of GMI, accounting for the dependent censoring of GMI. Extensive simulation studies show that our proposed estimators perform well in practical situations and outperform existing estimators. A phase II oncology trial is provided for illustration.

11:30-11:45 AM

Power Calculation for Cross-Sectional Stepped Wedge Cluster Randomized Trials with a Time-to-Event Endpoint

Mary Ryan, Yale School of Public Health; Yale Center for Analytical Sciences

Stepped wedge cluster randomized trials (SW-CRTs) are a form of randomized trial where clusters are progressively transitioned from control to intervention, and the timing of transition is randomized for each cluster. An important task at the design stage is to ensure that the planned trial has sufficient power to observe clinically meaningful effects. While methods for determining study power have been developed for SW-CRTs with continuous, binary or count outcomes, limited methods are available for SW-CRTs with censored time-to-event outcomes. We propose a stratified marginal Cox model to account for confounding by time in SW-CRTs, and develop analytical power calculation methods based on a nested Archimedean copula that differentiates between within- and between-period Kendall's Tau for the latent survival time. We also derive the explicit expression of the robust sandwich variance. Power formulas based on both the Wald and robust score tests are analytically developed and compared via simulations, demonstrating different finite-sample behaviors. Finally, we illustrate our methods using the context of a recently designed SW-CRT with a censored time-to-event outcome.

11:45-12:00 PM

Inference of Treatment Effect and its Regional Modifiers Using the Restricted Mean Survival Time in Clinical Trials Conducted in Multiple Regions

Kaiyuan Hua, Duke University

Multi-regional clinical trials (MRCTs) are playing an increasingly important role in pharmaceutical products developments by accelerating data gathering and regulatory approval to needed patients worldwide. The baseline characteristics of study participants often differ across regions of a single MRCT due to differences in recruitment practice or genetic and demographic profile. Unfortunately, existing methods on MRCTs have ignored the incomparability of baseline covariates, which may yield biased estimates of treatment effect and misleading conclusion on treatment effect consistency across regions. We develop a new weighting method that calibrates covariate distributions of the samples from different regions against to a given target population. We propose calibration weighting (CW) adjusted restricted mean survival times (RMSTs) and use RMST ratio to quantify the average treatment effects for the target population. We establish the consistency and asymptotic limiting distribution of the CW-adjusted RMSTs. Simulation studies are conducted to study the finite sample properties of the proposed estimator. We illustrate the proposed method using the data from a real MRCT.

12:00-12:15 PM

A Systematic Evaluation of Statistical Approaches for Non-Proportional Hazards

Xinyu Zhang, Yale School of Public Health

We conducted a systematic review of statistical methods designed for the time-to-event outcomes under various nonproportional hazard scenarios. Our study used data from published oncology trials to compare the Log-rank test against the MaxCombo test, the Restricted Mean Survival Time (RMST) test, the Generalized Gamma and Generalized F models. Power, type 1 error, and bias for RMST difference and survival probability difference were evaluated to compare the performance. Additionally, we constructed six scenarios with crossing hazards chosen so that the early and late effects ?cancel out? and used them to evaluate the ability to detect segment-specific and overall treatment effects. Recommendations, when each method performs the best, are given depending on the modeled scenario and the target question. In order to detect a treatment effect within a specific segment of the trial (time or information fraction), we proposed an approach to detect the change point in the Cox model by maximizing the partial likelihood function. Simulation results indicate the calculated change points are reasonably close to the true values.

20. CONTRIBUTED PAPERS: CAUSAL INFERENCE AND MEDIATION ANALYSIS

Chair: Haotian Zheng, University of Pennsylvania

10:30-10:45 AM

Semiparametric Bayesian Inference for Causal Mediation in Cluster Randomized Trials

Woojung Bae, University of Florida

We propose semiparametric Bayesian inference for causal mediation in cluster randomized trials (CRT). We estimate direct and indirect effects at the individual level and cluster level. For the cluster-specific confounder distributions, we specify a hierarchical Bayesian bootstrap (HBB) prior. This avoids restrictive parametric assumptions for confounder distribution at the individual level, and cluster level, and also enables us to borrow information across clusters. The observed data model along with causal assumptions allows us to identify and estimate the natural direct, and indirect effects at the individual level and cluster level. Simulation studies are presented to examine the performance of this approach.

10:45-11:00 AM

A Mediation-Based Analysis Approach for Assessing the Role of Engagement with Text Message-Delivered Interventions

Jamie Joseph, Vanderbilt University

Recent studies have utilized interactive text messages to support medication adherence in patients with conditions such as hypertension and diabetes. In general, it is understood that a text message-delivered intervention?s benefit should be derived at least in part by engagement with the intervention (quantifiable via, e.g., subject-specific response rate). Isolating the role of engagement as a mediator between an intervention and an outcome of interest is difficult, particularly due to the strong access monotonicity condition (i.e., subjects assigned to a control condition are unable to engage with the intervention). In this talk, we formalize the assumptions necessary to examine the role of engagement in the intervention?s effect using g-computation. Examples of target parameters under this framework include various direct and indirect effects, as traditionally identified in mediation analysis. Through simulation and application to a recent study of a texted-based intervention on HbA1c in adults with type 2 diabetes, we demonstrate that g-computation can be implemented in mobile health intervention settings in order to help identify effects of interest.

11:00-11:15 AM

Assessing Causal Mediation with Longitudinal Biomarkers Using Functional Regression Methods

David Cheng, Biostatistics Center, Massachusetts General Hospital

In clinical studies there is often interest in characterizing the degree to which effects of exposures or interventions are mediated through biomarkers. Traditional methods for mediation have assumed scalar mediators. However, biomarkers are dynamic processes in which shapes of trajectories over time may be informative beyond levels observed at any single time point or simple scalar summaries (e.g. peak values) of the trajectories. We consider longitudinal mediators as realizations of an underlying mediator process over possibly sparse and irregular time grids. We introduce semiparametric estimators for natural direct and indirect effects on outcomes by a fixed time mediated over the process based on inverse probability weighting, sequential regressions, and an augmented combination, leveraging functional regressions for the nuisance functions. Simulations demonstrate the reduced bias relative to methods relying on scalar trajectory summaries and the improved efficiency of the augmented estimator. We consider an application estimating the effects of obesity on severe disease mediated through inflammatory biomarkers among patients hospitalized with COVID-19.

11:15-11:30 AM

A Novel Causal Mediation Analysis Approach for Zero-Inflated Mediators

Meilin Jiang, University of Florida

Mediation analyses play important roles in making causal inference in biomedical research to examine causal pathways that may be mediated by one or more intermediate variables. Although mediation frameworks have been well established such as potential-outcomes models and traditional linear mediation models, little effort has been devoted to dealing with mediators with zero-inflated structures due to challenges associated with excessive zeros. We develop a novel mediation modeling approach to address zero-inflated count mediators containing true zeros and false zeros. The new approach can decompose the total mediation effect into two components induced by zero-inflated structures: the first component is attributable to the change in the mediator on its numerical scale which is a sum of two causal pathways and the second component is attributable only to its binary change from zero to a non-zero status. An extensive simulation study is conducted to assess the performance and it shows that the proposed approach outperforms existing standard causal mediation analysis approaches. We also showcase the application of the proposed approach to a real study.

11:30-11:45 AM

Robust Transfer Learning of Individualized Treatment Rules

Zhiyu Sui, University of Pittsburgh

Causality-based individualized treatment rules (ITRs) is a stepping stone to precision medicine. To ensure unconfoundedness, ITRs are ideally derived from randomized experimental data, but the use cases of ITRs in the real world extend far beyond these controlled settings. It is of great interest to transfer knowledge learned from experimental data to real world data but hurdles remain. In this paper, we address two challenges in the transfer learning of ITRs. 1) In well-designed experiments, granular information that is crucial to decision making can be thoroughly collected. However, part of this may not be accessible in real-world decision-making. 2) Experimental data with strict inclusion criteria reflects a population distribution that may be very different from the real-world population data, leading to biased estimation of ITRs. We propose a unified weighting scheme to learn a calibrated and robust ITR that simultaneously addresses the issues of covariate shift and covariate availability. The performance of this method is evaluated in simulation and real-data applications.

11:45-12:00 PM

High-Dimensional Mediation Analysis via Deep Neural Networks

Shuoyang Wang, Yale University

Mediation analysis draws increasing attention in many research areas such as epidemiology, genomics and psychology, and it gets difficult when the dimension of potential mediators is larger than the sample size. In addition, the effect of potential confounders can be complicated on both mediators and the outcome, which is merely considered as linear among existing works. In this paper, we propose a novel deep neural network estimation and inference procedure for evaluating the indirect and direct effect in mediation models, where linear high-dimensional mediators and non-linear confounders are incorporated in both the mediator model and the outcome model. By using the penalized method for partially linear regression via neural networks, the proposed procedure performs consistently in selecting active mediators and has strong prediction ability. Compared to existing methods, the superiority of the proposed approach is demonstrated in various Monte Carlo simulations and one real data application of childhood trauma.

12:00-12:15 PM

Federated and Transfer Learning Approaches for Causal Inference

Larry Han, Harvard University

It is challenging to accurately estimate treatment effects for underrepresented populations. Data fusion can improve power, but privacy concerns often constrain the sharing of patient-level data. We propose a framework to leverage multiple sites and multiple populations to make inference on the conditional average treatment effect (CATE) for an underrepresented target population of interest. Our method leverages transfer and federated learning to data-adaptively incorporate summary-level information from source populations and sites to learn about treatment effects for underrepresented populations. When it is undesirable to define subgroups a priori, we propose an alternative federated causal tree approach. In extensive simulation studies, we show that the proposed methods substantially improve the estimation accuracy of treatment effects for underrepresented target populations, lowering RMSE by up to 80%. We illustrate our method through a real-world study of COVID-19 vaccine efficacy on infection, hospitalization, and mortality in underrepresented populations using data from multiple hospitals.

21. CONTRIBUTED PAPERS: CLUSTERED DATA METHODS

Chair: Samuel Anyaso-Samuel, University of Florida

10:30-10:45 AM

Game Theory Based Functional/Longitudinal Data Clustering

Xiang Wang, IUPUI

Game theory based clustering is a sequential partitioning of data points to maximize the within cluster similarity using the dominant set concept from graph theory, which is different from common existing methods such as K-means clustering, hierarchical clustering and spectral clustering. We propose a hierarchical bipartition procedure under the penalized optimization framework with the tuning parameter selected by maximizing modularity of the resulting two clusters. The proposed game theory based hierarchical method is applied to longitudinal/functional data clustering with a flexible choice of similarity measures between curves. It is not only robust to uneven sizes of clusters but also to outliers, which overcomes the limitation of many existing clustering methods. We demonstrate the benefits of the proposed method via extensive empirical investigations using simulations as well as real data applications.

10:45-11:00 AM

Health Care Provider Clustering Using Fusion Penalty in Quasi-Likelihood

Lili Liu, Washington University in St. Louis

There has been growing research interest in developing methodology to evaluate the health care providers' performance with respect to a patient outcome. Random and fixed effects models are traditionally used for such a purpose. We propose a new method, using fusion penalty to cluster health care providers based on quasi-likelihood. Without any priori knowledge of grouping information, our method provides a desirable data-driven approach for automatically clustering health care providers into different groups based on their performance. Further, the quasi-likelihood is more flexible and robust than the regular likelihood in that no distributional assumption is needed. An efficient alternating direction method of multipliers algorithm is developed to implement the proposed method. We show that the proposed method enjoys the oracle properties; namely, it performs as well as if the true group structure were known in advance. The consistency and asymptotic normality of the estimators are established. Simulation studies and analysis of the national kidney transplant registry data demonstrate the utility and validity of our method.

11:00-11:15 AM

Bayesian Keep-or-Merge Training Framework for Data Integration in ERP-based Brain-Computer Interface

Tianwen Ma, Emory University

An event-related potential (ERP)-based BCI speller helps disabled people with normal communications. Existing methods constructed binary classifiers to detect target ERP responses. Current training strategy uses data from participants themselves only with lengthy training time, which causes attention shifts and mental fatigue. To resolve this issue, we propose a Bayesian Keep-or-Merge (BKM) method for data integration. BKM specifies the joint distribution of stimulus-specific EEG signals among new and source participants via a Bayesian hierarchical mixture model. We refer to the baseline cluster as the one for the new participant. For inference, we apply a keep-or-merge strategy such that if source and new participants are similar, they share the same set of model parameters, otherwise, they keep their own sets of model parameters. The similarity is determined by a binary selection indicator vector. The parameter set for source participants can be computed ahead of time. For prediction, we predict on the testing data with the baseline cluster directly. We demonstrate the advantages of BKM using extensive simulation studies and show the real data analysis from XXX Lab.

11:15-11:30 AM

A Hypothesis Test for the Detection of Spatial Clustering in Areal Data

Stella Watson, University of South Carolina

Spatial clustering detection has many applications, such as identifying disease outbreaks or crime hotspots. Ripley?s K function is a common method for detecting clustering in point process data, measuring the expected number of events within a given distance of any observed event. While performing spatial clustering analysis on point process data is common, applications to areal data are also of interest. For example, researchers might wish to determine if tracts of conserved land are clustering together to create larger preserved areas. In this work, we develop a new function for quantifying clustering in areal data inspired by Ripley?s K function. We then use this function to develop a hypothesis testing procedure to detect clustering and/or dispersion in areal data at specific distances. We compare the performance of our method to other cluster detection methods, including the spatial scan, global Moran?s I and general Getis-Ord G statistics. Finally, we employ our new method to an actual dataset of conservation easements in Boulder County, Colorado to determine if the conservation easement tracts are spatially clustering.

11:30-11:45 AM

Pursuing Sources of Heterogeneity in Microbiome Community Structure via a Regularized Dirichlet-Multinomial Mixture Model

Zhongmao Liu, University of Connecticut

In microbiome studies, it is often of great interest to identify natural clusters or partitions in the samples and characterize the unique compositional profile for each microbial community. Different metacommunities can differ in taxonomic, functional, ecological and medical properties which can help with the evaluation of human health and diseases. Built upon the Dirichlet multinomial mixture (DMM) model, we propose a novel sparse DMM method with group sparsity to simultaneously detect sample clusters and identify sources of heterogeneity, i.e., the critical taxa which differentiate metacommunities. A comprehensive simulation study shows that the sparse DMM method achieves good feature selection performance in identifying heterogeneous taxa. An application to the upper-airway microbiota and asthma study on children shows that nasal microbiome groups defined by the sparse DMM method have differential risk levels of loss of asthma control in the future.

11:45-12:00 PM

Sparse Clusterability: Testing for Cluster Structure in High Dimensions

Naomi Brownstein, Medical University of South Carolina

Cluster analysis is commonly used to separate data into groups. For cluster analysis to be meaningful, the data to be clustered is assumed to be generated from a population consisting of multiple distinct clusters. Clusterability analysis facilitates testing of the inherent assumption of latent cluster structure. This presentation highlights methods for clusterability testing with high-dimensional data. Proposed methods combine sparse principal component analysis with tests of multimodality. Type I error and power of the clusterability tests are presented on simulated data with different types of cluster structure. The methods are then applied to real-world data in the area of gene expression, microarray, and shotgun proteomics. To our knowledge, our work is the first analysis of clusterability testing for high-dimensional data.

12:00-12:15 PM

Assessing Treatment Effect Heterogeneity in the Presence of Missing Effect Modifier Data in Cluster-Randomized Trials

Bryan Blette, University of Pennsylvania

The assessment of heterogeneous treatment effects (HTE) based on pre-specified potential effect modifiers has become a common goal in randomized trials. However, when effect modifiers are missing, complete-case analysis may lead to bias, under-coverage, inflated type I error, or low power. While statistical methods have been proposed and compared for individually randomized trials with missing effect modifier data, few guidelines exist for the cluster-randomized setting, where intracluster correlations may introduce further issues in assessing HTE. In this work, we propose a Bayesian multilevel multiple imputation (MMI) method and compare its performance to complete case analysis, single imputation, multiple imputation, and standard MMI in a simulation study of cluster-randomized trials with missing effect modifier data. The methods are further compared using real data from the Work, Family, and Health Study. The results suggest that MMI and Bayesian MMI have the best overall performance and that

Bayesian MMI has improved bias and coverage over MMI when there are model specification or compatibility issues.

22. CONTRIBUTED PAPERS: DIAGNOSTIC AND SCREENING TESTS

Chair: Adam Ciarleglio, The George Washington University

10:30-10:45 AM

Comparing Net Benefit-Risk for Diagnostic Tests and Biomarkers Under Tree Ordering

Jing Kersey, Georgia Southern University

The evaluation or comparison of diagnostics tests and biomarkers based on benefit-risk involves both the accuracy of the tests and the clinical consequences of the diagnostic errors. In practice, many diseases can be classified into multiple classes. Besides monotone ordering, another important category of scenarios for multi-classes is tree ordering, in which the diseases consist of several unordered subclasses. In diseases with multi-subclasses without an order, the benefits and risks of the clinical consequences could differ from class to class. Investigations on the benefit-risk of diagnostic tests with more than two classes are lacking in the literature. This paper extends the diagnostic yield table in binary disease cases to clinical conditions with multi-subclasses. Moreover, a decision process based on net benefit for evaluating diagnostic tests is developed. The proposed decision process provides additional interpretation for rule-in or rule-out clinical conditions and their adverse consequences from unnecessary workups in diseases with multi-subclasses. Simulations and real data examples are presented to illustrate the proposed measures.

10:45-11:00 AM

Medical Diagnostics Accuracy Measures and Cut-Point Selection: An Innovative Approach Based on Relative Net Benefit

Hani Samawi, Georgia Southern University

For some diseases the prevalence is either unknown or different from region to region or population to population, resulting in an erroneous diagnosis. This paper introduces innovative post-test diagnostic accuracy measures and a new cut-point selection criterion based on the expected relative net benefit. Our approach does not depend on the disease's prevalence, maximizing net benefit and reducing the clinical consequences of diagnostic errors. We demonstrate the advantages of the proposed measures to compare different diagnostic tests and/or biomarkers, on average, the abilities for rule-in, rule-out clinical conditions, and cut-point selection criteria that maximize the expected relative net-benefit diagnostic accuracy. Numerical examples, simulation studies, and real data are provided to illustrate the superiority and applicability of the proposed measures.

11:00-11:15 AM

A Threshold-Free Summary Index of Prediction Accuracy

Qian Zhou, Mississippi State University

Positive predictive values (PPV) have been recommended for evaluating the performance of a risk-scoring system to predict the risk of having an event by a prespecified future time. However, for a continuous or ordinal risk score, the PPV requires a subjective cutoff threshold value that dichotomizes the score, which creates barriers for practitioners and researchers. We proposed a threshold-free summary index of PPV: the average precision (AP). When evaluating the performance change between an existing risk model and a new one, we often find that the incremental value (IncV) in AP does not always agree with the IncV in other metrics, such as AUC. This disagreement can create confusion when assessing whether the added information improves the model prediction accuracy. We have discovered the analytical connections and differences between the AUC IncV and AP IncV. These two IncV metrics are both weighted averages of the changes in separating the risk score distributions between events and non-events. However, AP IncV assigns heavier weights to the changes in higher-risk regions, whereas AUC IncV weights the changes equally.

11:15-11:30 AM

A Collapsing Net Benefit Approach to Evaluating Alzheimer's Disease Biomarkers Using Benefit-Risk Measures

Ferdous Ahmed, Levine Cancer Institute, Atrium Health

Clinicians may be skeptical of comparing diagnostic tests or biomarkers based solely on accuracy measures. Diagnostic test accuracy is measured in clinical settings using classification accuracy or predictive values. The drawback of these metrics is one test may have higher sensitivity but worse specificity than another. Another approach is to compare tests or biomarkers using a benefit-risk measure, which entails quantifying test benefits and clinical consequences of diagnostic errors. Diagnostic tests are commonly classified into positive and negative categories (diseased or non-diseased). Some diseases, such as Alzheimer's, have more than two stages. The benefit to cost values may fluctuate depending on the stage of the disease. This study demonstrates the application of the net benefit approach to evaluating biomarkers for the diagnosis of Alzheimer's disease (AD) based on a clinical performance study and clinical effects. As a result, we present a diagnostic yield table and develop a decision theory based on net benefit for analyzing biomarkers, which provides interpretation for rule-in or rule-out clinical conditions.

11:30-11:45 AM

Evaluating Diagnostic Test Accuracy in Studies with Seemingly Fatal Verification Bias

Gene Pennello, U.S. Food & Drug Administration, Center for Devices and Radiological Health

Diagnostic tests are evaluated for accuracy based on agreement of the test result with the reference standard that verifies true disease status (absent, present). Missing reference standard results can preclude evaluation of some of the common test accuracy measures such as sensitivity, specificity, diagnostic likelihood ratio (negative, positive), ROC, NPV, or PPV. For example, for rule-in tests, data are sometimes available only on PPV and the probability of a test positive because subjects are referred to the reference for verification of disease status only if they test positive. With these data, we show how to estimate a1 = (PPV ? p) / (1 ? p), where p is the prevalence of disease, i.e., its pre-test probability. Likewise, for rule-out tests with data available only on NPV and the probability of a test positive, we show how to estimate a0 = (p ? (1 ? NPV)) / p, which is called the attributable risk in epidemiology. Note that a0 and a1 standardize the differences PPV ? p and p ? (1 ? NPV), which are the pre-test to post-test disease probability stratifications.

11:45-12:00 PM

Performance of Diagnostic Tests Based on Continuous Bivariate Markers

Marwan Alsharman, Georgia Southern University

Collecting multiple continuous biomarker measures is customary to improve the accuracy of diagnostic tests. A prevalent practice is combining these biomarkers' measurements into one single composite score. However, incorporating those biomarker measurements into a single score depends on the combination of methods and may lose vital information needed to make an accurate decision. Furthermore, a diagnostic cut-off is required for such a combined score, which is difficult to interpret in clinical practice. The paper extends the classical biomarkers? accuracy and predictive values from univariate to bivariate markers. Also, we will develop a novel pseudo-measures system to maximize the vital information from multiple biomarkers. We specified these pseudo-and-or classifiers for the true positive rate, true negative rate, false-positive rate, and false-negative rate. We used them to redefine classical measures such as the Youden index, diagnostics odds ratio, likelihood ratios, and predictive values. We provide optimal cut-off point selection based on the modified Youden index with numerical illustrations and real data analysis.

12:00-12:15 PM

Efficient Risk-Based Collection of Biospecimens in Cohort Studies: Application to Diagnostic Performance of Multicancer Early Detection Tests

Mark Ramos, National Cancer Institute

In cohort studies, it is of interest to subsample from the full cohort when biospecimen collection is too expensive to do on the entire cohort. We focus on collecting biospecimens to estimate sensitivity of diagnostic tests like Multi-Cancer Early Detection (MCED) assays. We propose a three-phase sampling design to enrich the number of cases for different diseases, given available risk models for each disease, in order to oversample people most likely to become cases. An adaptive estimator was developed to tradeoff bias vs. variance given unknown dependence between diagnostic test status and each risk model. Results from simulations and application to the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial data show that these designs perform at least as well as simple random sampling, and that relative performance improves dramatically as we have more predictive risk models, smaller subsample sizes, or increased dependence between risk status and diagnostic test status within disease status.

23. CONTRIBUTED PAPERS: ENVIRONMENTAL AND ECOLOGICAL APPLICATIONS

Chair: Jimmy Hickey, North Carolina State

10:30-10:45 AM

A Family of Partial Linear Single Index Distributed Models for Analyzing Environmental Mixtures and Implementation in R Package EPLSIM

Yuyan Wang, NYU Grossman School of Medicine

Statistical methods to study the joint effects of environmental mixtures are of great importance to understanding the impact of correlated exposures that may act synergistically or antagonistically on health outcomes. This study proposes a family of statistical models under a unified partial-linear single-index (PLSI) modeling framework, to assess the joint effects of environmental factors for continuous, categorical, time-to-event, and longitudinal outcomes. All PLSI models consist of a linear combination of exposures into a single index for practical interpretability of relative direction and importance, and a nonparametric link function for modeling flexibility. These models were demonstrated using a dataset of 800 subjects from the NHANES 2003?2004 survey. We have developed an R package named “EPLSIM” for Environmental analysis using this PLSI family and will present implementations of the above models using the package.

10:45-11:00 AM

An Overview of Methods for Environmental Mixture Exposures Subject to Detection Limits

Myeonggyun Lee, National Institute of Environmental Health Sciences

In environmental epidemiology, it is of great interest to understand the impact of environmental mixtures on human health. Even though there are popular modeling approaches such as weighted quantile sum regression (WQS) and Bayesian kernel machine regression (BKMR), studies on the effects of environmental mixtures face the challenge of limit of detection (LOD) in multiple correlated exposure measurements. Several methods have been proposed to incorporate exposures subject to LOD in risk modeling using standard regression models. However, these methods have not been investigated under WQS, BKMR and other popular approaches. In this study, we performed extensive simulations to evaluate the performance of existing methods for environmental mixtures with different approaches for covariates subject to LOD. We considered (i) complete case analysis, (ii) fill-in by LOD divided by the square root of two, (iii) multiple imputation, and (iv) fill-in by conditional expectation using a parametric accelerated failure time (AFT) model. Our simulation study and data application were illustrated using National Health and Nutrition Examination Survey (NHANES) dataset.

11:00-11:15 AM

Penalized Dynamic Single-Index Models

Yiwei Li, NYU Langone Health Department of Population Health Division of Biostatistics

There is substantial interest in assessing the time-dependent environmental mixtures? effects on human health. Existing methods mainly focus on linear exposure-outcome relationships and thus inadequately address the nonlinear dependencies between exposure mixtures and outcomes of interest. We therefore propose a novel penalized dynamic single-index model to study the accumulative effect of multiple time-dependent exposures on a scalar outcome. The effects of exposures are aggregated by a single index combination. Two types of constraints are used to describe the synergistic or antagonistic exposure effects. P-spline approach is adopted to approximate the unknown bivariate link function, using penalized regression splines with roughness penalties can approximate any flexible function well and avoid overfitting. An iterative estimation algorithm is proposed. The large sample properties of the estimator are established, and the performance of the method is evaluated under extensive simulation scenarios. The application of our method is demonstrated by a cohort study examining the effect of prenatal environmental pollutant exposures on fetal birth weight.

11:15-11:30 AM

Influence of Post-Traumatic Stress and Abnormal Spirometry on Cognitive Performance in 9/11 WTC Responders

Jaeun Choi, Albert Einstein College of Medicine

Post-traumatic stress disorder (PTSD) and abnormal spirometry are highly prevalent mental and health conditions in World Trade Center (WTC) responders. We hypothesized that PTSD symptomatology and abnormal spirometry are synergistically associated with cognitive performance in WTC responders. PTSD symptomatology was assessed using the PCL-IV, and we calculated the FEV1/FVC ratio to measure pulmonary function and characterize abnormal spirometry. Cogstate assessment measured cognitive performance. We evaluated PTSD, pulmonary function and their interaction on cognitive performance by linear regressions adjusting for confounders. PTSD symptomatology and pulmonary function appeared to have a significant synergistic effect on cognitive performance in that higher severity of PTSD symptomatology in the presence of lower pulmonary function was associated with poorer cognitive performance. Results suggested chronic stress and lung damage might share underlying biological mechanisms, including inflammatory and oxidative stress pathways, which might also affect the brain. Early intervention efforts to mitigate preventable cognitive decline in high-risk populations should be studied.

11:30-11:45 AM

The Impact of Geomasking on Estimating Health Effects of Spatial Environmental Exposures

Kayleigh Keller, Colorado State University

Protecting participant privacy is paramount for cohort studies with detailed individual-level data. But ensuring privacy can be challenging when studies also wish to make data publicly available. This challenge is particularly acute when geographic location of study participants is of scientific interest, such as when evaluating the health effects of environmental exposures that vary spatiotemporally. Geomasking provides an approach to preserving privacy of participants by perturbing their true locations. However this process introduces measurement error that can have a deleterious effect on epidemiological inference. In this work, we evaluate the tradeoff between participant privacy and accurate estimation of an exposure-response relationship for long-term air pollution exposures via simulation and in a large US cohort.

11:45-12:00 PM

Did the COVID-19 Lockdowns Improve Air Quality? Machine-Learning Based Robust Estimation of Effects of Policy Interventions on Air Pollution

Claire Heffernan, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

Studies examining effects of policy interventions (such as COVID-19 lockdowns) or other natural experiments on air pollutant concentrations are abundant. However, careful statistical analysis is important to deconvolve causal changes from other contributing factors. We present a comprehensive framework for studying causal changes due to such perturbations that guards against false conclusions. We outline the assumptions required for identifying the causal changes and show that many common methods often implicitly relied on strong and unrealistic assumptions. We present flexible machine learning-based interrupted time series models that produce a causal estimand and do not use restrictive assumptions. For validation, we propose a strategy of guarding against falsely significant effects in past years when there was no intervention. The framework is applied to study the impact of COVID-19 lockdowns on NO2 concentrations in the eastern US, which reveals frequent false conclusions by commonly used methods but not by machine learning approaches. We find significant decreases in NO2 concentrations due to COVID-19 lockdowns in Boston, New York City, Baltimore, and Washington D.C.

12:00-12:15 PM WITHDRAWN

Trends in Air Pollution: A Quantile Regression Approach

Sukanya Bhattacharyya, North Carolina State University

Concern regarding climate change and its influential impact on humanity is the talk of the hour. Air pollutant levels in air are constantly monitored, and we use the United States Environmental Protection Agency's available resources to access the distribution of particular pollutants for a given number of sites, over the years. Various spatial locations have their spatially dependent pollutant?s quantile functions which varies with time. Using an approach of simultaneously modelling the quantiles, our aim is to reduce the computational complexity than the existing methodologies. We use a quantile regression method that uses functional principal components to reduce the dimensions over space and quantile levels while testing for trends in air pollution data over the last 20 years. Extensive comparison among the existing methods in literature is demonstrated.

24. CONTRIBUTED PAPERS: EPIDEMIOLOGICAL METHODS

Chair: Theresa Devasia, NIH

10:30-10:45 AM

A Flexible Zero-Inflated Conway--Maxwell--Poisson Regression Model for Spatiotemporal Data of US Vaccine Refusal

Bokgyeong Kang, The Pennsylvania State University

Vaccination is widely acknowledged as one of the most effective tools for preventing disease. However, there has been a rise in parental refusal and delay of childhood vaccination in recent years in the United States. This trend undermines the maintenance of herd immunity and elevates the likelihood of outbreaks of vaccine-preventable diseases. Our aim is to identify demographic or socioeconomic characteristics associated with vaccine refusal, which could help public health professionals and medical providers develop interventions targeted to concerned parents. We examine US county-level vaccine refusal data for patients under five years of age collected on a monthly basis during the period 2012--2015. These data exhibit challenging features: zero inflation, spatial dependence, seasonal variation, spatially-varying dispersion, and a large sample size (approximately 3,000 counties per month). We propose a flexible zero-inflated Conway--Maxwell--Poisson (ZICOMP) regression model that addresses these challenges. The class of ZICOMP models contains an intractable normalizing constant, which makes Bayesian inference challenging. We propose a new hybrid Monte Carlo algorithm that permits efficient sampling and provides asymptotically exact estimates of model parameters.

10:45-11:00 AM

Improving Association Analysis of Self-Reported Outcome with a Validation Subset: An Application to the Childhood Cancer Survivor Study

Sedigheh Mirzaei Salehabadi, St. Jude Children's Research Hospital

In health-science research, outcomes ascertained through surveys are subject to potential bias with respect to true outcomes, which are only ascertainable with clinical assessment. This measurement error can lead to biased inferences on association measures between outcomes and exposures of interest. Here we consider a cohort study in which the outcome of interest is subject to imperfect ascertainment, a common situation with questionnaire-based outcome ascertainment. But, a clinically assessed outcome is available for a subset of the cohort. This presents an opportunity to address the bias challenges. Specifically, we construct a likelihood consisting of one from the validation subset and the other from the non-validation subset. Pepe (1992) offered one previously, and the other proposed here is motivated to enable inference with standard statistical software. MLEs are computed and statistical inference is based on the standard large-sample likelihood theory. This method was motivated by a long-term childhood cancer survivors cohort study. The application example is discussed.

11:00-11:15 AM

Conditional or Unconditional Logistic Regression for Matched Case-Control Studies?

Fei Wan, Washington University in St Louis

Although the need for addressing matching in the analysis of matched case-control studies is well established, debate remains as to the most appropriate analytical method. To address these uncertainties, we viewed matched case-control matching from the perspective of weighted sampling and derived the outcome models describing how the exposure and matching factors are associated with the outcome in the individual and frequency matched designs. In either case the derived outcome model is a logit model with stratum-specific intercepts. We compared the bias and efficiency of unadjusted and adjusted conditional logistic regression (CLR) and unconditional logistic regression (ULR) for both individually and frequency matched case-control designs. In both designs, ULR is more vulnerable to model specification error and is generally biased unless the functional form of the matched factors is modeled correctly. CLR should remain the primary analytical approach for applied researchers.

11:15-11:30 AM

Causal Decomposition Maps: An Exploratory Tool for Designing Area-level Interventions Aimed at Reducing Health Disparities

Melissa Smith, University of Alabama at Birmingham

Methods for decomposition analyses have been developed to partition between-group differences into explained and unexplained portions. We introduce the concept of causal decomposition maps, which allow researchers to test the effect of area-level interventions on disease maps before implementation. These maps quantify the impact of interventions that aim to reduce differences in health outcomes between groups. They also illustrate how the disease map might change under different interventions. We adapt a causal decomposition analysis method for the disease mapping context. Through the specification of a Bayesian hierarchical outcome model that allows for spatial interference of the intervention, we obtain counterfactual small area estimates of age-adjusted rates and reliable estimates of decomposition quantities. We highlight our method?s use for designing area-level interventions aimed at reducing rural-urban differences in age-adjusted cancer rates in Iowa ZIP codes.

11:30-11:45 AM

Spatial Confounding Overstates Protective Effect of COVID Vaccine at County Level, 2021

Pavel Chernyavskiy, University of Virginia School of Medicine

To investigate the protective effects of COVID vaccination on health outcomes (e.g., COVID infection, hospitalization, mortality), we often use spatial regression models with vaccination data taken at some geographic unit, such as state or county. However, this approach leads to high potential for spatial confounding between the vaccination data and the spatial random effect, precluding an accurate estimate. Here, we use the newly-developed spatial+ approach (Dupont et al., 2022) applied to county level data to generate an unbiased estimate in the presence of confounding. Prior to adjusting for spatial confounding, the protective effect of the COVID vaccine was 9% (95% CrI 6%, 11%) per 15% of population vaccinated. After adjusting for spatial confounding, he protective effects ranged from 6.9% to 8.5%, depending on model used. Results motivate the need for caution when analyzing COVID data: failing to adjust for spatial confounding can overstate protective effects by as much as 30%.

11:45-12:00 PM

Comparing the Degree of Oversmoothing in Methods for Disease Mapping

Jihyeon Kwon, Drexel University

In the context of disease mapping ? where the goal is to make inference on incidence and/or prevalence rates of diseases in small areas ? recent work has developed an approximation for the informativeness of the popular conditional autoregressive (CAR) model framework and highlighted its tendency to overwhelm the contribution of the data, thereby producing overly smooth and overly precise estimates. The objective of the current study is to assess the degree to which several recently developed, alternative models for disease mapping suffer from the same afflictions. After deriving expressions for the informativeness of the various models, we analyze data comprised of the number of heart disease related deaths in Pennsylvania counties and demonstrate the degree to which all of the models considered produce rate estimates with a similar degree of oversmoothing. We then illustrate how restrictions can be imposed on each of the models to reduce the informativeness of the models to a desired level.

12:00-12:15 PM

Fractal Dimension Based Geographical Clustering of COVID-19 Time Series Data

Yessika Adelwin Natalia, Data Science Institute, Hasselt University

Understanding the local dynamics of COVID-19 transmission calls for an approach that characterizes the incidence curve in a small geographical unit. Given that incidence curves exhibit considerable day-to-day variation, the fractal structure of the time series dynamics is investigated for the Flanders and Brussels Regions of Belgium. For each statistical sector, the smallest administrative geographical entity in Belgium, fractal dimensions of COVID-19 incidence rates, based on rolling time spans of 7, 14, and 21 days were estimated using four different estimators: box-count, Hall-Wood, variogram, and madogram. We found varying patterns of fractal dimensions across time and location. The fractal dimension is further summarized by its mean, variance, and autocorrelation over time. These summary statistics are then used to cluster regions with different incidence rate patterns using k-means clustering. Fractal dimension analysis of COVID-19 incidence thus offers important insight into the past, current, and arguably future evolution of an infectious disease outbreak.

Monday, March 20, 2023 | 1:45-3:30 PM

25. ADVANCES IN STATISTICAL METHODS FOR CELL-TYPE-SPECIFIC ANALYSES

Organizer/Chair: Hongyu Zhao, Yale University

1:45-2:10 PM

Robust and Accurate Estimation of Cellular Fractions from Tissue Omics Data via Ensemble Deconvolution

Jiebiao Wang, University of Pittsburgh

Dozens of cellular deconvolution methods have been proposed to infer cellular fractions from bulk omics data; however, these methods are sensitive to real application settings. Benchmarking showed no uniformly best deconvolution approaches. There are two existing ensemble methods, but they only aggregate multiple single-cell references or reference-free methods. To achieve a robust estimation of cellular fractions, we proposed EnsDeconv (Ensemble Deconvolution), which adopts cell type-specific (CTS) robust regression to synthesize the results from dozens of single deconvolution methods, reference datasets, marker gene selection procedures, data normalizations, and transformations. Unlike most benchmarking based on simulations, we compiled four large real datasets of 4937 bulk samples with measured cellular fractions and bulk gene expression from different tissue types. Comprehensive evaluations demonstrated that EnsDeconv yields more stable, robust, and accurate fractions than existing methods, enabling various CTS downstream analyses, such as differential fractions associated with clinical variables. EnsDeconv was further extended to analyze bulk DNA methylation data.

2:10-2:35 PM

Cell-Type-Specific Co-Expression Inference with Single-Cell RNA Sequencing Data

Emma Zhang, University of Miami

Gene co-expression networks inferred from microarray and sequencing data offer valuable information on the functional organization of genes. The advancement of single cell RNA-sequencing technology enables researchers to study co-expression networks at the individual cell type level. However, the high noise and heterogeneity of single cell data present great challenges in recovering true expression levels from the observed counts, and lack of attention to these issues in existing single cell network inference methods may lead to biased estimates, inflated type I error and reduced power. In this talk, we describe CS-CORE, a statistical method that is built on an expression-measurement model tailored to single cell data to explicitly model the technical noises. We show that CS-CORE can decouple co-expression from measurement noises for single cell co-expression estimation and testing in both simulations and real data.

2:35-3:00 PM

Dozer: Debiased Personalized Gene Co-Expression Networks for Population-Scale scRNA-Seq Data

Sunduz Keles, University of Wisconsin - Madison

Population-scale single cell RNA-seq (scRNA-seq) datasets create unique opportunities for quantifying expression variation across individuals at the gene co-expression network level. Gene-gene correlation estimates from scRNA-seq tend to be severely biased towards zero for genes with low and sparse expression. We present Dozer to debias correlation estimates from scRNA-seq datasets and quantify network level variation across individuals. Dozer corrects correlation estimates in the general Poisson measurement model and provides a metric to quantify genes measured with high noise. Computational experiments establish that Dozer estimates are robust to mean expression levels of the genes and the sequencing depths of the datasets. Compared to alternatives, Dozer results in fewer false positive edges in the co-expression networks, yields more accurate estimates of network centrality measures and modules, and improves the faithfulness of networks estimated from separate batches of the datasets. We showcase unique analyses enabled by Dozer in two population-scale scRNA-seq applications.

3:00-3:25 PM

Cell-Type-Specific Gene Regulation in Human Neural Progenitors and Neuronal Progeny

Michael Love, University of North Carolina at Chapel Hill

Cell-type-specific analysis of gene regulation can help reveal tissues and time points for which genetic liability for disease is conferred. In collaboration with a team of scientists at the UNC Neuroscience Center, we have performed cell-type-specific accessibility, expression, and splicing quantitative trait locus (QTL) analysis, as well as allelic analysis, on two key cell types implicated in neuropsychiatric disorders. I will discuss the opportunities such datasets provide for identifying mediating biomarkers for disease, and provide some examples where cell-type-specific QTL and allelic results colocalize with brain-trait related GWAS signals. In addition, I will discuss statistical methods for detecting cell-type-specific gene regulation across cell sub-populations via newly developed single cell allelic expression assays.

26. BRAIN FUNCTIONAL CONNECTOME IN FMRI STUDIES

Organizer/Chair: Yi Zhao, Indiana University

1:45-2:10 PM

Statistical Harmonization Methods for Multi-Site Functional MRI Studies

Russell Shinohara, University of Pennsylvania

As multi-center functional magnetic resonance imaging studies have been come commonplace, pooling and integrating data from multi-site studies has become critical. Site differences attributed to various sources are known to exist and might result in a substantial impact on the analytic results. Recently, batch-effect correction methods such as ComBat and CovBat have been successfully adapted to remove scanner and site differences in functional neuroimaging data and applied in many large-scale studies. However, fewer methods are available to harmonize the resting-state functional magnetic resonance imaging (fMRI) connectivity matrices, given the complex dependency structures temporally and spatially in the raw fMRI data and that the derived connectivity matrices do not necessarily belong to Euclidean metric space. We will discuss several extensions of the statistical harmonization methods to multisite functional connectivity data and white matter hyperintensity data.

2:10-2:35 PM

Counteracting Selection Bias in Functional Connectivity Studies of Autism

Benjamin Risk, Emory University

In resting-state functional magnetic resonance imaging studies, it is common for more than 50% of data to be removed due to participant motion. Motion tends to be higher in children with developmental disorders, such as autism spectrum disorder. This creates the potential for selection bias. In this study, we modify doubly robust estimators of the average treatment effect to address this problem. We propose a permutation test of the difference between two groups for improved type one error rates in finite samples. We compare functional connectivity in autistic children to children without autism.

2:35-3:00 PM

Mapping the Connectome: Statistical Learning for Reliable Brain Network Analysis

Ying Guo, Emory University

Brain network-oriented analyses have become increasingly popular in neuroimaging studies to advance understanding of neural circuits and their association with neurodevelopment, mental illnesses and aging. These analyses often encounter challenges such as low signal-to-noise ratio in neuroimaging data, the high dimensionality of brain networks, and the large number of brain connections leading to spurious findings. In this talk, we present several new statistical methods to improve reliability and reproducibility in whole-brain connectomics. The proposed methods tackle the aforementioned challenges by multimodality integrative network modeling to investigate the interplay between white matter structural connection and functional connection using a multilevel probabilistic model, and by providing robust blind source separation of observed brain connectivity matrices to reveal underlying neural circuits and their association with demographic and clinical factors. We will discuss the theoretical properties and computational advantages of the methods, and demonstrate their performance through simulation studies and real-world neuroimaging data examples.

3:00-3:25 PM

Thresholded Prior for Integrating Brain Regional and Network Predictors

Yize Zhao, Yale University

The investigation of intrinsic connectivity networks by resting-state fMRI has proven capable of revealing fundamental elements of human brain architecture and organization. Optimal integration of data from different fMRI modalities is an active area of research aimed at increasing diagnostic accuracy. On the other side, how to select features from high-dimensional measures as biomarkers for building a model to predict brain physiology is an important and challenging problem. We build a Bayesian model to predict cognitive behavior from multimodality brain imaging data, specifically working memory fMRI and resting-state connectivity fMRI, and identify potential biomarkers by imposing a jointly sparse structure. By rank-R PARAFAC decomposition of the coefficient matrix for the network predictor, we could achieve subnetwork topography simultaneously. We apply our method to the Adolescent Brain Cognitive Development (ABCD) Study by identifying associated region-level and subnetwork-level neuroimaging biomarkers and evaluate our methods by extensive simulations.

27. INFORMATIVE PRIOR ELICITATION FOR COMPLEX INNOVATIVE CLINICAL TRIALS

Organizer: Joseph Ibrahim, University of North Carolina at Chapel Hill
Chair: Xinxin Chen, University of North Carolina at Chapel Hill

1:45-2:10 PM

The Latent Exchangeability Prior (LEAP) for Borrowing Information from Historical Controls in Clinical Trials

Ethan Alt, University of North Carolina at Chapel Hill

In clinical trials, it is often desirable to borrow information from historical controls. For example, if a study utilizes two-to-one randomization, borrowing information from historical controls can help increase efficiency in estimating a treatment effect. However, some individuals in the historical study may be more relevant than others at explaining the outcome. The current state of the art utilizes propensity score methods, where the probability of belonging to the current study is modeled as a function of covariates. Notably, this approach does not borrow information based on the outcome, which could lead to bias. In this talk, we introduce the latent exchangeability prior (LEAP). The LEAP uses latent classes to explicitly model whether a subject in the historical data is exchangeable as those in the current data. Unlike propensity score approaches, the LEAP propagates uncertainty over who (if anyone) is exchangeable with subjects in the current data set. We compare our approach using simulations and a real data analysis to other popular priors.

2:10-2:35 PM

Prior Elicitation and Pediatric/Adolescent Trial Design Evaluation in Cases with Anticipated Efficacy Attenuation Compared to Adults

Matthew Psioda, GSK

Extrapolating efficacy from adult to younger populations is a valuable strategy to address the challenge of producing substantial evidence of treatment benefit in those settings. As a result, prior elicitation is a critical component of study design. In this talk we propose a novel three-component robust mixture prior (RMP) which includes a robustification component, a bridging component, and a prior reflecting the effect as characterized by the adult trial data. The bridging component represents belief that the effect in adolescents may be attenuated relative to that in adults and is informed using two Bayesian meta-regression models based on summary statistics from other studies in in the same disease populations. To obtain the bridging prior, the adult program posterior is modified by rescaling its mean and standard deviation using results from the Bayesian meta-regression models. Thus, the informative RMP component reflects a compromise between the effect observed in adults for the same product and an attenuated effect based on extrapolating the attenuation factor from other, similar programs.

2:35-3:00 PM

The Scale Transformed Power Prior for Use with Historical Data from a Different Outcome Model

Joseph Ibrahim, University of North Carolina

We develop the scale transformed power prior for settings where historical and current data involve different data types, such as binary and continuous data, respectively. This situation arises often in clinical trials, for example, when historical data involve binary responses and the current data involve time-to-event or some other type of continuous or discrete outcome. The power prior proposed by Ibrahim and Chen (2000) does not address the issue of different data types. Herein, we develop a new type of power prior, which we call the scale transformed power prior (straPP). The straPP is constructed by transforming the power prior for the historical data by rescaling the parameter using a function of the Fisher information matrices for the historical and current data models, thereby shifting the scale of the parameter vector from that of the historical to that of the current data. Examples are presented to motivate the need for a scale transformation and simulation studies are presented to illustrate the performance advantages of the straPP over the power prior and other informative and non-informative priors.

3:00-3:25 PM

DISCUSSANT

Ram Tiwari, Bristol Myers Squibb

28. SYNTHETIC CONTROL METHODS UNDER INTERFERENCE

Organizer: Taylor Krajewski, University of North Carolina at Chapel Hill
Chair: Michael Hudgens, University of North Carolina at Chapel Hill

1:45-2:10 PM

The Inclusive Synthetic Control Method

Giovanni Mellace, University of Southern Denmark

We introduce the inclusive synthetic control method (iSCM), a modification of synthetic control type methods that allows the inclusion of units potentially affected directly or indirectly by an intervention in the donor pool. This method is well suited for applications with either multiple treated units or in which some of the units in the donor pool might be affected by spillover effects. Our iSCM is very easy to implement using most synthetic control type estimators. As an illustrative empirical example, we re-estimate the causal effect of German reunification on GDP per capita allowing for spillover effects from West Germany to Austria.

2:10-2:35 PM

Direct and Spillover Effects Using Synthetic Control and Matrix Completion Methods

Giulio Grossi, University of Florence

In recent years, Synthetic Control (SC) and Matrix Completion (MC) methods are becoming increasingly popular among scholars interested in drawing causal claims. Such methods rely on non-interference assumption, which could be unreasonable in particular settings, such as spatial or network data structures. We investigate the use of the SC and MC methods in panel comparative case studies where interference between the treated and the untreated units is plausible. We frame our discussion in the potential outcomes approach. Under a partial interference assumption, we define relevant direct and spillover effects. We address the spatial structures of the data, underlying the main differences between the available methods. Then we investigate the assumptions under which we can identify and estimate the causal effects of interest, and show how they can be estimated using the SC and MC methods. We apply our approach to the analysis of an observational study, where the focus is on assessing direct and spillover causal effects of a new light rail line recently built in Florence (Italy)

2:35-3:00 PM

Estimating the Effectiveness of Permanent Price Reductions for Competing Products Using Multivariate Bayesian Structural Time Series Models

Fiammetta Menchetti, DiSIA, University of Florence

The Florence branch of an Italian supermarket chain recently implemented a strategy that permanently lowered the price of numerous store brands in several product categories. To quantify the impact of such a policy change, researchers often use synthetic control methods. In our application, however, competitor brands not assigned to treatment are likely impacted by the intervention because of substitution effects. This paper extends synthetic control methods to allow interference within predefined groups but not between them. Focusing on a class of causal estimands capturing the effect both on the treated and control units, we develop a multivariate Bayesian structural time series model for generating synthetic controls that would have occurred in the absence of an intervention. In a simulation study we explore our Bayesian procedures' empirical properties and show that it achieves good frequentists coverage, even when the model is misspecified. We use our new methodology to make causal statements about the impact on sales of the affected store brands and their direct competitors. Our proposed approach is implemented in the CausalMBSTS R package

3:00-3:25 PM

Estimation and Inference for Synthetic Control Methods with Spillover Effects

Jianfei Cao, Northeastern University

The synthetic control method is often used in treatment effect estimation with panel data where only a few units are treated and a small number of post-treatment periods are available. Current estimation and inference procedures for synthetic control methods do not allow for the existence of spillover effects, which are plausible in many applications. In this paper, we consider estimation and inference for synthetic control methods, allowing for spillover effects. We propose estimators for both direct treatment effects and spillover effects and show they are asymptotically unbiased. In addition, we propose an inferential procedure and show it is asymptotically unbiased. Our estimation and inference procedure applies to cases with multiple treated units or periods, and where the underlying factor model is either stationary or cointegrated. In simulations, we confirm that the presence of spillovers renders current methods biased and have distorted sizes, whereas our methods yield properly sized tests and retain reasonable power.

29. VANDERBILT BIOSTATISTICS AFTER 20 YEARS: LESSONS LEARNED AND OPPORTUNITIES AHEAD FOR A BIOSTATISTICS DEPARTMENT IN A MEDICAL CENTER

Organizer/Chair: Bryan Shepherd, Vanderbilt University Medical Center

1:45-2:10 PM

My Big Jump: Founding a Department of Biostatistics

Frank Harrell, Vanderbilt University School of Medicine

For many years biostatistics had been successful at Vanderbilt, but the opportunity to create a department home for biostatistics was too good to pass up. The new department and its support from the School of Medicine leadership made it an attractive place for recruiting new faculty and staff. This talk will cover what made the department attractive, as well as principles upon which the Department of Biostatistics was founded in the Vanderbilt School of Medicine in 2003. These principles include reproducible research and prioritizing collaboration over consultation. Challenges and opportunities of running the department in a growing academic medical center will be discussed, with emphasis on generalizable knowledge that may assist others in starting, sustaining, and enhancing biostatistics groups in their own medical centers.

2:10-2:35 PM

Growing Biostatistics and Data Science Education Programs: The Vanderbilt Experience

Qingxia Chen, Vanderbilt University Medical Center

Vanderbilt's Biostatistics Graduate Program was founded about a decade ago. As a small program established in a medical center during the big data era, we have worked to expand the skills and knowledge of a new generation of statisticians beyond traditional statistical modeling and inference. I will first share how we developed a course curriculum that emphasizes reproducible research and encourages interdisciplinary collaboration while maintaining a strong methodological identity. Unlike traditional programs supported mainly by tuition and/or state revenue, we rely heavily on external funding. Although this poses significant challenges, it's also a tremendous opportunity for our students, exposing them to real-world biomedical research requirements early in their academic careers. I will also discuss how we worked to attract talent and motivate students to join our new, small program, how we're developing a distance learning program, and how we're collaborating and competing with other departments/schools in the data science education niche.

2:35-3:00 PM

What Keeps Me Awake at Night Leading a Department of Biostatistics

Yu Shyr, Department of Biostatistics Vanderbilt University Medical Center

The emergence of big data has advanced the goals of precision medicine; however, across the entire continuum of big data capture and utilization, many more challenges lie ahead—from analysis of high-throughput biomarkers to maximum exploitation of the electronic health record. Because of these challenges, the statistics profession is in a period of disruptive change that John Tukey foresaw almost 60 years ago. He pointed to the existence of an as-yet unrecognized science in his book, The Future of Data Analysis. Today, if the statistical community does not participate in the data revolution, we will be marginalized; if we do not adapt our mindset, we will find ourselves relegated to a supporting role on the data science stage; if we do not educate our students on new concepts in statistics, we will be less and less successful in passing the statistical torch. In this presentation, I will offer some perspectives on the changing landscape for statistical science; the need for statisticians to adjust their mindset around the explosive growth in information technology; machine learning; and the AI revolution.

3:00-3:25 PM

DISCUSSANT

David L. DeMets, University of Wisconsin-Madison

30. BAYESIAN MACHINE LEARNING FOR DECISION-MAKING WITH INCOMPLETE INFORMATION

Organizer: Joe Hogan
Chair: Taylor Fortnam

1:45-2:10 PM

Causal Framework for Subgroup Treatment Evaluation Using Multivariate Generalized Mixed Effect Models with Longitudinal Data

Yizhen Xu, Johns Hopkins University

Dynamic prediction of causal effects under different treatment regimes conditional on individual's characteristics and longitudinal history is an essential problem in precision medicine. One of the challenges is that the existence of selection bias is empirically untestable. We propose a framework for identifying the long-term individualized treatment effect adjusting for unobserved stable trait factors, using Bayesian G-computation with multivariate generalized mixed effect models. Existing methods mostly focus on balancing the confounder distributions of observables between different treatments, while our proposal also accounts for the latent tendency towards each treatment due to unobserved time-invariant factors. We assume sequential ignorability conditional on unobserved stable trait factor in treatment assignment, and dynamically updates stable unobserved factors in outcomes progression as an individual's history data increases over time. Our framework naturally incorporates sensitivity analysis, providing an alternative to defining an additional sensitivity parameter for quantifying the impact of unmeasured confounding.

2:10-2:35 PM

Bayesian Semiparametric Model for Sequential AML Treatment Decisions with Informative Timing

Arman Oganisian, Brown University

We develop a Bayesian semiparametric model for the impact of dynamic treatment rules (DTRs) on survival among patients diagnosed with pediatric acute myeloid leukemia (AML). The data are from a phase III clinical trial in which patients move through a sequence of four treatment courses. At each course, they undergo chemotherapy that may or may not include anthracyclines (ACT). While ACT is known to be effective, it is also cardiotoxic and can lead to early death. Estimation of potential survival probability under hypothetical DTRs is challenging for several reasons. First, since ACT was not randomized in the trial, its effect on survival is confounded over time. Second, the timing of the next course depends on recovery from the previous course - making timing potentially informative. Third, patients may die or drop out before ever completing the full treatment sequence. We use Gamma Processes to model the underlying continuous-time transition process between subsequent death and treatment states. A g-computation procedure is used to compute posterior potential survival probabilities under various DTRs that tailor ACT inclusion based on evolving cardiac function.

2:35-3:00 PM

Bayesian Machine Learning for Causal Inference with Incomplete Longitudinal Covariates and Censored Survival Outcomes

Jungang Zou, Columbia University

For missing at random longitudinal covariates, existing imputation methods are primarily based on parametric models, in which exact relationships among longitudinal response, treatment, and covariates are explicit. Misspecification of the parametric form can introduce model misspecification biases. We propose a nonparametric Bayesian sequential imputation method for missing at random longitudinal covariates. We first develop a novel nonparametric Bayesian trees mixed-effects model to model the longitudinal trajectories flexibly. We then develop an efficient MCMC algorithm to sequentially impute the missing longitudinal covariates data. The novel longitudinal missing data methodology is then formally integrated with a robust survival g-formula to study the effect of longitudinal treatment on patient survival. We conduct extensive simulations in the context of longitudinal treatments and incomplete longitudinal covariates to investigate the practical operating characteristics of our proposed methods. Finally, we apply our methods to a pooled dataset from six NHLBI cohorts to estimate the optimal dynamic antihypertensive treatment initiating rules for young or older adults.

3:00-3:25 PM

DISCUSSANT

Jason Roy, Rutgers University

31. CONTRIBUTED PAPERS: SPARSE DATA MODELING AND METHODS

Chair: Ying Cui, Emory University

1:45-2:00 PM

Robust Two-Layer Partition Clustering of Sparse Multivariate Functional Data

Zhuo Qu, King Abdullah University of Science and Technology

In this work, a novel elastic time distance for sparse multivariate functional data is proposed. Subsequently, a robust distance-based two-layer partition clustering is introduced. With the proposed distance, our approach not only can detect correct clusters for sparse multivariate functional data under outlier settings but also can detect those outliers that do not belong to any clusters. The classical distance-based clustering methods such as density-based spatial clustering of applications with noise (DBSCAN), agglomerative hierarchical clustering and K-medoids are extended to the sparse multivariate functional case based on our proposed distance. Numerical experiments on the simulated data highlight that the performance of the proposed algorithm is superior to the performances of the existing model-based and extended distance-based methods. Using Northwest Pacific cyclone track data as an example, we demonstrate the effectiveness of the proposed approach. The code is available href{https://github.com/ZhuoQu/Sparse_multivariate_functional_clustering}{online} for readers to apply our clustering method and replicate our analyses.

2:00-2:15 PM

Bayesian Generalized Linear Low Rank Regression Models for the Detection of Vaccine-Adverse Event Associations

Paloma Hauser, University of North Carolina at Chapel Hill

We propose a generalized linear low-rank mixed model (GLLRM) for the analysis of both high-dimensional and sparse responses and covariates where the responses may be binary, counts, or continuous. This development is motivated by the problem of identifying vaccine-adverse event (AE) associations in post-market drug safety databases. The GLLRM is a generalization of a generalized linear mixed model (GLMM) in that it integrates a factor analysis model to describe the dependence among responses and a low rank matrix to approximate the high-dimensional regression coefficient matrix. A sampling procedure combining the Gibbs sampler and Metropolis and Gamerman algorithms is employed to obtain posterior estimates of the regression coefficients and other model parameters. Testing of response-covariate pair associations is based on the posterior distribution of the corresponding regression coefficients. Monte Carlo simulation studies are conducted to examine the finite-sample performance of the proposed procedures on binary and count outcomes. We further illustrate the GLLRM via a real data example based on the Vaccine Adverse Event Reporting System (VAERS).

2:15-2:30 PM

Spike-and-Slab LASSO Additive Cox Model for High-dimensional Survival Prediction And Functional Selection

Boyi Guo, Johns Hopkins University Bloomberg School of Public Health

Cox proportional hazards (PH) model is one of the most popular models in biomedical data analysis. There have been continuing efforts to improve the flexibility of such models for complex signal detection, for example, via spline functions. Nevertheless, it is nontrivial to extend to the high-dimensional setting (p>>n). When estimating spline functions, commonly used grouped sparse regularization may induce excess shrinkage, damaging the predictive performance. Moreover, the previous “all-in-all-out” strategy for functional selection fails to answer if nonlinear components exist. We develop an additive Cox PH model that employs a novel spike-and-slab LASSO prior to select the linear and nonlinear components of spline functions. A scalable and deterministic algorithm, EM-Coordinate Descent, is designed for efficient model fitting. We compare the predictive and computational performance against the state-of-the-art models via Monte Carlo studies and metabolomics data analysis. The proposed model is broadly applicable to various research fields, e.g., genomics and population health, via the freely available R package BHAM.

2:30-2:45 PM

Spatial Predictions on Physically Constrained Domains: Applications to Arctic Sea Salinity Data

Bora Jin*, Duke University

We predict sea surface salinity (SSS) in the Arctic Ocean based on satellite measurements. SSS is a crucial indicator for ongoing changes in the Arctic Ocean and can offer important insights about climate change. We particularly focus on areas of water mistakenly flagged as ice by satellite algorithms. To remove bias in the retrieval of salinity near sea ice, the algorithms use conservative ice masks, which result in considerable loss of data. We aim to produce realistic SSS values for such regions to obtain more complete understanding about the SSS surface over the Arctic Ocean. We propose a class of scalable nonstationary processes that can handle large data from satellite products and complex geometries of the Arctic Ocean. Barrier Overlap-Removal Acyclic directed graph GP (BORA-GP) constructs sparse directed acyclic graphs (DAGs) with neighbors conforming to barriers and boundaries, enabling characterization of dependence in constrained domains. The BORA-GP models produce more sensible SSS values in regions without satellite measurements and show improved performance in various constrained domains in simulation studies compared to state-of-the-art alternatives.

2:45-3:00 PM

Sparse High-Dimensional Linear Mixed Modeling with a Partitioned Empirical Bayes ECM Algorithm

Anja Zgodic, University of South Carolina

High-dimensional longitudinal data is increasingly used in a wide range of biological studies. However, statistical methods for high-dimensional linear mixed models (LMMs) remain few, as most Bayesian variable selection or penalization methods are designed for independent observations. In this work, we present an efficient and accurate Bayesian framework for high-dimensional LMMs. This method offers increased flexibility through the development of empirical Bayes estimators for hyperparameters, with computationally efficient estimation through the Expectation Conditional-Maximization (ECM) algorithm. The novelty of the approach lies in its partitioning and parameter expansion, which allows maximum a posteriori probability (MAP) estimation of parameters. We illustrate Linear Mixed Modeling with PaRtitiOned empirical Bayes ECM (LMM-PROBE) in simulation studies evaluating the estimation of fixed and random effects, as well as variance components. A real-world example is provided using the riboflavin dataset, where we identify genes associated with the decline of riboflavin production in recombinant Bacillus subtilis bacteria and predict production rate over time.

3:00-3:15 PM

Variable Selection for Competing Risks in High-Dimensional Covariate Spaces with Missing Data

Guowei Li, The Ohio State University

Competing risks data sometimes arise in the clinical setting when the primary event of interest competes with one or possibly several other events. The goal is to model the time to the primary event of interest, for example, death due to a specific cause, using available predictors. In high-throughput genomic studies, the number of features often far exceeds the number of subjects, thus it is challenging to select a parsimonious set of features that predicts the outcome. To further complicate the issue, some predictors might have missing values. Here, we propose a variable selection method based on the penalized proportional subdistribution hazards model defined on multiple imputed datasets. Using simulation studies, we show that this method works well in high-dimensional settings and generally selects the important predictors and few unimportant predictors. We demonstrate this method when modeling time-to-relapse for acute myeloid leukemia patients who have achieved complete remission using demographic, clinical, and genomic features.

32. CONTRIBUTED PAPERS: ESTIMATION AND PREDICTION UNDER MISSINGNESS/MISCLASSIFICATION

Chair: Zeling He, Emory University

1:45-2:00 PM

Statistical Inference for Association Studies in the Presence of Binary Outcome Misclassification

Kimberly Hochstedler*, Cornell University

In biomedical and public health association studies, a binary outcome variable may be subject to misclassification, resulting in substantial bias in effect estimates. The feasibility of addressing binary outcome misclassification in regression models is often hindered by model identifiability issues. In this paper, we characterize the identifiability problems in this class of models as a specific case of "label switching" and leverage a pattern in the resulting parameter estimates to solve the permutation invariance of the complete data log-likelihood. Our proposed algorithm for correcting label switching in binary outcome misclassification models relies only on the assumption that outcomes are correctly classified at least 50% of the time and does not require gold standard labels. This label switching correction is applied within estimation methods to recover unbiased effect estimates and to estimate misclassification rates. Open source software is provided to implement the proposed methods. We give a detailed simulation study for our proposed methodology and apply these methods to data from the Medical Expenditure Panel Survey (MEPS) from 2020.

2:00-2:15 PM

Estimating Marginal Treatment Effect in Cluster Randomized Trials with Multi-level Missing Outcomes

Chia-Rui (Jerry) Chang*, Harvard University T.H. Chan School of Public Health

Cluster randomized trials (CRTs), where clusters of individuals are randomized to different intervention conditions, are commonly used in biomedical research, but their analyses can be complicated by missing outcomes. When outcome missingness depends on baseline covariates, methods such as IPW-GEE can be used to estimate the marginal treatment effect. These methods are developed to handle the setting where missingness occurs at the individual level and each cluster has partially or fully observed individual outcomes. When all outcomes from a cluster are missing, these approaches ignore this cluster-level missingness and can lead to biased inference. We propose a new estimator, multi-level multiply robust GEE (MMR-GEE), to account for multi-level missingness in estimating the marginal treatment effect, which allows analysts to specify multiple sets of PS models. The proposed estimator is consistent and asymptotically normal provided that one set of the models is correctly specified. We evaluate the performance of the proposed method through extensive simulations and illustrate its use with a CRT evaluating a Malaria risk-reduction intervention in Ghana.

2:15-2:30 PM

Mining for Equitable Health: Assessing the Impact of Missing Data in Electronic Health Records

Emily Getzen, University of Pennsylvania/School of Medicine

Electronic health records (EHRs) contain years of health information to be leveraged for precision health. However, they can present significant analytical challenges? they contain multi-scale data that are collected at irregular time intervals and with varying frequencies. In addition to these challenges, EHRs can reflect inequity-- the data for marginalized groups may be less informative due to more fragmented care, which can be viewed as a missing data problem. For EHR data in this complex form, there is currently no framework for introducing missing values. There has also been little to no work in assessing the impact of missing data in EHRs. In this work, we simulate realistic missing data scenarios in EHRs to adequately assess their impact on predictive modeling. We incorporate the use of a medical knowledge graph to capture dependencies between medical events to create a more realistic framework. In an intensive care unit setting, we found that missing data have greater negative impact on the performance of disease prediction models in groups that tend to have less access to care.

2:30-2:45 PM

Submodel Approximation for Risk Prediction of a New Patient with Missing Risk Factors

Tianyi Sun, Vanderbilt University

Clinical prediction models have been widely acknowledged as an informative tool that provides evidence-based support for clinical decision making. However, such prediction models are often underused in clinical practice due to many reasons including the presence of missing information in a new patient. Motivated by a study to implement a prediction model STRATIFY into the clinical workflow of emergency department (ED), we propose a novel submodel approach to address real-time missing information issues. For prediction models such as STRATIFY that were developed using preconditioning outcome, the proposed submodel coefficients are shown to be equivalent to the original prediction model coefficients plus a corrected factor. Comprehensive simulations were conducted to assess the performance of the proposed estimation approach and compared with an existing one-step-sweep approach using various performance measures. The proposed approach was applied to electronic health records data from the ED at Vanderbilt University Medical Center to develop submodels for STRATIFY which will subsequently be embedded in the STRATIFY clinical decision support tool for real-time implementation.

2:45-3:00 PM

Mission Imputable: Robustly Modeling Huntington Disease Progression in the Presence of Covariate Censoring

Kyle Grosser, University of North Carolina at Chapel Hill

To select potent outcomes for trials of Huntington disease, a fatal disorder, analysts first model how possible outcomes change over time using data from untreated subjects. Yet subjects are often observed at different disease stages. To account for this, analysts include time of diagnosis as a covariate; however, this covariate is often censored. With many solutions to covariate censoring, we impute censored values using predictions from a model of the covariate, then analyze the imputed dataset. Yet when our covariate model is wrong, bias can ensue. To remedy this, we first model imputed diagnosis times as error-prone versions of the true diagnosis times. We then adjust for this error using semiparametric theory to derive an estimating equation that estimates our outcome model given censored covariates, even when those covariates are modeled incorrectly. We show that our method 1) produces identifiable and consistent model estimates, 2) remains empirically unbiased unlike competing methods which yield >100 bias, and 3) pinpoints potent outcomes for clinical trials aimed at slowing Huntington disease progression, even when applied to heavily censored data.

3:00-3:15 PM

Testing Unit Root Non-Stationarity in the Presence of Missing Data in Univariate Time Series of Mobile Health Studies

Charlotte Fowler, Columbia University Mailman School of Public Health, Department of Biostatistics

The use of digital devices to collect data in mobile health (mHealth) studies introduces a novel application of time series methods, with the constraint of potential data missing at random (MAR) or missing not at random (MNAR). In time series analysis, testing for stationarity is an important preliminary step to inform appropriate later analyses. The augmented Dickey-Fuller (ADF) test was developed to test the null hypothesis of unit root non-stationarity, under no missing data. Existing methods for time series with missing data such as complete case analysis, last observation carry forward, multiple imputation with chained equations, and linear interpolation impose constraints on the autocorrelation structure, and thus impact unit root testing. We propose maximum likelihood estimation and multiple imputation using a state space model approaches to adapt the ADF test to a context with missing data. We further develop sensitivity analysis techniques to examine the impact of MNAR data. We evaluate the performance of existing and proposed methods across different missing mechanisms in extensive simulations and in their application to a multi-year smartphone study of bipolar patients.

3:15-3:30 PM

Transportable Risk Prediction with Heterogeneous Electronic Health Records

Guanghao Zhang, Department of Biostatistics, University of Michigan

Medical codes across electronic health record (EHR) systems and across coding systems are usually heterogeneous, which hinders the development of transportable algorithms and affects the performance when algorithms are directly applied to a heterogeneous EHR source. Most transfer learning methods ignore heterogeneity in different data sources and existing methods for dealing with heterogeneity across EHR sources require manual collation and are therefore not scalable. Assuming heterogeneity across different EHR sources is determined by an orthogonal mapping matrix, we propose estimation methods of the mapping matrix such that features from these two heterogeneous EHR sources can be automatically mapped. This scalable data-driven method accommodates both one-to-one and one-to-many mapping patterns. With the estimated mapping matrix, we provide a model transfer algorithm that transfers model fitted with the training data to the heterogeneous target data and present a computationally efficient proposition for transferring generalized linear model. We demonstrate the validity of our proposed method through simulation studies and an application study with real-world data.

33. CONTRIBUTED PAPERS: REGRESSION MODELS AND METHODS FOR FUNCTIONAL DATA

Chair: Xin Ma, Florida State University

1:45-2:00 PM

Nonparametric Functional Data Modeling of Pharmacokinetic Processes with Applications in Dynamic PET Imaging

Baoyi Shi, Columbia University Mailman School of Public Health

Modeling a pharmacokinetic process typically involves solving a system of linear differential equations and estimating the parameters upon which the functions depend. In order for this approach to be valid, it is necessary that a number of fairly strong assumptions hold, assumptions involving various aspects of the kinetic behavior of the substance being studied. In many situations, such models are understood to be simplifications of the "true" kinetic process. While in some circumstances such a simplified model may be a useful (and close) approximation to the truth, in some cases, important aspects of the kinetic behavior cannot be represented. We present a nonparametric approach, based on principles of functional data analysis, to modeling of pharmacokinetic data. We illustrate its use through application to data from a dynamic PET imaging study of the human brain.

2:00-2:15 PM

Ensemble Classification Using Shape-based Distance for Dynamic PET Imaging Data

Denise Shieh, Columbia University

Positron Emission Tomography (PET), a nuclear medicine imaging technique, is an invaluable tool to study a broad range of mental illnesses as it enables quantitative measurements of the density of various proteins throughout the brain. A PET scan begins with the injection of a radioactive substance, a radiotracer, that has affinity for the protein of interest. The density of the target protein at each location can then be inferred from the concentration data by modeling the kinetic behavior of the radiotracer. The impulse response function (IRF) provides the key to understanding the kinetic behavior of the tracer, and thus all PET modeling methods involve calculating estimates of this function. Existing methods summarize the entire IRF with a single scalar measure. Our goal is to utilize distance metrics based on principles of functional data analysis (FDA) and shape data analysis for comparison of subjects and take a k-nearest-neighbor ensemble approach to optimally combine the metrics.

2:15-2:30 PM

A Bayesian Scalar-On-Function Quantile Regression with Measurement Error Using the GAL Distribution

Roger Zoh, Indiana University

Quantile regression provides a consistent approach to investigating the association between covariates and various aspects of the distribution of the response. Most methods assume that the covariates in the regression are precisely measured with no potential for errors. Here, we propose extending the Bayesian measurement error and Bayesian quantile regression literature to allow for functional covariates prone to measurement error. Our approach uses the Generalized Asymmetric Laplace (GAL) distribution. The family of Gal distribution has recently emerged as a more flexible distribution family in Bayesian quantile regression modeling. We compare the proposed approach's performance to that of a naive approach that ignores measurement error. We finally apply our approach to the analysis of an NHANES dataset.

2:30-2:45 PM

A Multivariate Functional Mixed Effects Model for Longitudinal Functional Data with Application in Biomechanical Data from 300 Recreational Runners

Edward Gunning, University of Limerick

We have developed a novel multivariate functional mixed effects model for multilevel longitudinal functional data. Our work is based on a study investigating the relationship between running-related injuries and biomechanical variables in 300 recreational runners. Kinematic motion analysis data (i.e., joint angles) were collected for multiple joints (hip, knee, ankle) on both legs while the participants ran on a treadmill for 3 minutes. The dataset consists of almost 50,000 multivariate functional observations and has an intricate multilevel structure ? multiple strides from the same individual are measured on both sides of the body. The strides also have a natural time ordering that needs to be considered when developing a model.

The multivariate functional mixed effects model allows us to quantify the effect of previous injury and other covariates on the multivariate functional data while accounting for variability between subjects, within subject between sides (i.e., asymmetries) and from stride-to-stride. We also show that including the time-ordering of the strides in the model leads to improved predictions at an individual level.

2:45-3:00 PM

Simultaneous Clustering and Decomposition of Neural Activation Data Across Repeated Trials

Madison Stoms, Columbia University

A better understanding of the neural activation patterns that emerge during voluntary skilled movements can provide insights into the connection between the motor cortex and observed behavioral outcomes. In the experiment that motivates our work, high-resolution firing rates from a collection of neurons were collected as a trained mouse made repeated reaches for a food pellet under multiple experimental conditions. Our goal is to group neurons in a data-driven way that leverages the repeated trials while allowing for interpretable cluster-and-condition-specific representations of the data. We extend methods from functional data analysis ? a collection of methods which considers data measured over a continuous domain ? for the data we observe and the goals of our analysis, in particular drawing on functional principal components analysis for dimension reduction and clustering. We illustrate our new methods in simulations and apply them to data collected on 25 neurons across 196 trials of two types.

3:00-3:15 PM

Shapes in the Mirror: Contrasting Waves of COVID-19 Epidemics through Functional Data Analysis

Jacopo Di Iorio, Pennsylvania State University

We use data from 107 Italian provinces to characterize and compare mortality patterns in the first two COVID-19 epidemic waves, which occurred prior to the introduction of vaccines. We associate these patterns with mobility, timing of government restrictions, and socio-demographic, infrastructural, and environmental covariates. Notwithstanding limitations in accuracy and reliability of publicly available data, we are able to exploit information in curves and shapes through Functional Data Analysis techniques. Specifically, we document differences in magnitude and variability between the two waves; while both were characterized by two mortality patterns (an "exponential" and a "milder" one), the second spread much more broadly and asynchronously through the country. Moreover, we find evidence of a significant positive association between local mobility and mortality in both epidemic waves, and corroborate the effectiveness of timely mobility restrictions in curbing mortality. Additional signals could be identified analyzing cases and positivity rates. However, as we shown, the quality of such data was too poor to support meaningful analyses.

3:15-3:30 PM

Generalized Additive Models of Ambulatory Blood Pressure Profiles

Raphiel Murden, Emory University - Rollins School of Public Health

Cardiovascular disease is the leading cause of death in women in the United States, and it disproportionately impacts Black women. Recent studies have shown that traditional risk factors, such as body mass index and smoking, do not explain this disparity and that psychosocial risk factors such as stress may be more informative. We examine the relationship between household financial responsibility and ambulatory blood pressure (ABP) among Black women in a longitudinal study conducted in metro Atlanta. Many previous studies of ABP profiles use either a summary measure or inflexible parametric models. However, these approaches may result in the loss of substantial variability or unnecessarily constrain profile shape. To overcome these limitations, we use generalized additive mixed models (GAMMs) with penalized cyclic splines to estimate average ABP profiles for women primarily responsible for earning household finances versus those who are not. We find that GAMMs enable us to assess periods of the day during which the groups differ significantly, which may inform interventions that help prevent adverse cardiovascular events.

34. CONTRIBUTED PAPERS: METHODS FOR SPATIOTEMPORAL-VARYING COVARIATES AND ESTIMATES

Chair: Kaidi Kang, Vanderbilt University

1:45-2:00 PM

Causal Exposure-Response Curve Estimation Using Generalized Propensity Score Matching in Cohorts with Geographically-Varying Study Eligibility Thresholds

Jenny Lee, Department of Biostatistics, Harvard T.H. Chan School of Public Health

Observational data often has multi-level structure where units are nested in clusters. However, one of the key challenges of the multi-level observational data is that there may exist unmeasured unit-level and cluster-level confounders due to lack of resources or due to privacy. To our knowledge, there is no literature on estimation of causal exposure-response function (ERF) in multi-level data. In this paper, we propose a matching method to estimate causal ERF with continuous exposure in multi-level data using a cluster adjusted generalized propensity score (cluster-GPS). A cluster-GPS is a weighted average of the distance between GPS values and a cluster-level values that enables matching units that belong to similar cluster using data-driven weight. We explain possible extensions of existing methods in multi-level data to ERF estimation and conduct simulation studies to evaluate the performance our proposed method relative to these approaches. We apply our proposed method to estimate the average causal ERF between long-term fine particulate matter exposure and respiratory hospitalization among low-income children in Medicaid during the period 2000-2012.

2:00-2:15 PM

Joint Modeling with Integrated Fractional Brownian Motion for Monitoring Disease Progression and Mortality

Anushka Palipana, Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center

It is difficult to characterize complex variations of biological processes, which are commonly measured using biomarkers that yield noisy data. While joint modeling with a longitudinal submodel for the biomarker measurements and a survival submodel for assessing the hazard of events can alleviate measurement error issues, the longitudinal submodel often uses random intercepts and slopes to estimate between-/within-patient heterogeneity in biomarker trajectories. To overcome longitudinal submodel challenges, we replace random slopes with scaled integrated fractional Brownian motion (IFBM). Using the IFBM submodel, we derive novel target functions to predict clinically important changes in the underlying longitudinal process to monitor the risk of rapid disease progression as real-time predictive probabilities. We use a Cox model for event hazard estimation. We implement the joint model in two stages via Bayesian posterior computation and inference using a patient registry and describe lung disease progression and mortality in lymphangioleiomyomatosis patients. IFBM, integrated Ornstein-Uhlenbeck, and conventional modeling are compared. Simulations are conducted.

2:15-2:30 PM

Clustering Multivariate Longitudinal Data Using Multidimensional Tensor Product Smoothing Splines and Composite Variables

Sherry Livingston, Medical University of South Carolina

Clinical research datasets often contain repeated measures of multiple biomarkers (e.g., lab values) of interest. These data are collected at differing frequencies, irregular intervals, can have high variability, and often have a correlation between variables that needs to be considered in statistical analysis. Some existing methods fail to take this correlation into account and can suffer drawbacks including long computation times and convergence issues. In this paper we present a method that combines multivariate tensor product smoothing splines with fuzzy clustering methods to create a balanced dataset with reduced noise that is then examined using a robust fuzzy clustering algorithm to identify clusters of individuals with similar trajectory shapes. We also investigate the use of a composite score calculated from multiple collected biomarkers. The approach is compared to other smoothing approaches in a simulation study and applied to a clinical setting. Our method results in better clustering accuracy than the use of individually smoothed variables, and the composite approach can offer a simpler implementation and interpretation in some instances.

2:30-2:45 PM

Maximum Likelihood Estimation for Semiparametric Regression Models with Interval-Censored Multi-State Data

Yu Gu, University of North Carolina at Chapel Hill

Interval-censored multi-state data arise in many studies of chronic diseases, where the health status of a subject can be characterized by a finite number of disease states and the transition between any two states is only known to occur over a broad time interval. We formulate the effects of potentially time-dependent covariates on multi-state processes through semiparametric proportional intensity models with random effects. We adopt nonparametric maximum likelihood estimation (NPMLE) under general interval censoring and develop a stable expectation-maximization (EM) algorithm. We show that the resulting parameter estimators are consistent and that the finite-dimensional components are asymptotically normal with a covariance matrix that attains the semiparametric efficiency bound and can be consistently estimated through profile likelihood. In addition, we demonstrate through extensive simulation studies that the proposed numerical and inferential procedures perform well in realistic settings. Finally, we provide an application to a major epidemiologic cohort study.

2:45-3:00 PM

A New Cure Model Accounting for Longitudinal Data and Flexible Patterns of Hazard Ratios Over Time

Can Xie, The University of Texas Health Science Center at Houston

With the advancement of medical treatments, many historically incurable diseases have become curable. Accurate estimation of the cure rates is of great interest. When there is no clear biomarker indicator for cure, the estimation of cure rate is intertwined and influenced by the specification of hazard functions for uncured patients. Consequently, the commonly used proportional hazard assumption, when it is false, may lead to biased cure rate estimation. To avoid this problem, we propose a new cure model with flexible hazard ratios between different covariate subgroups. It is a joint model with individual random effects shared between cure-survival and longitudinal submodels. The regression parameters are estimated by maximization of the non-parametric likelihood via the expectation-maximization algorithm, with an adaptive Metropolis algorithm used in the E steps. Simulation studies consider both crossing and non-crossing survival curves. In both situations, the proposed model provides unbiased estimates for the cure rates. Its application to a study of chronic myeloid leukemia demonstrates improved performance over widely used models on cure rate estimation.

3:00-3:15 PM

Inference for Signals Exhibiting Irregular Statistical Cyclicity with Applications to Electrocardiograms

Bartosz Majewski, AGH University of Science and Technology

In this paper, a bootstrap approach for the electrocardiogram (ECG) model based on an amplitude-modulated time-warped periodically correlated process is proposed. Specifically, the circular version of the Extension of Moving Block Bootstrap is tailored to handle the complex features of ECG data. The proposed method gives bootstrap pointwise and simultaneous confidence intervals for the real and the imaginary parts of the Fourier coefficients of the mean and the autocovariance functions of the underlying periodically correlated process.

The possible applications of the bootstrap inference for the ECG data are discussed. The example of real data analysis is presented. Moreover, based on the estimated underlying PC process, a new amplitude-modulated time-warped periodically correlated signal is simulated to examine the effectiveness of the discussed method. The confidence intervals are constructed based on the new signal and the results with the original ECG signal are compared.

35. CONTRIBUTED PAPERS: STATISTICAL METHODS FOR OMICS DATA

Chair: Linghua Wang, University of Texas MD Anderson Cancer Center

1:45-2:00 PM

Linkreg: A Bayesian Framework for Linking Candidate Cis-Regulatory Elements to Target Genes

Qiuhai Zeng, Pennsylvania State University

Cis-regulatory elements (CREs) are non-coding DNA segments that regulate transcription. Many candidate CREs (cCREs) have been identified in the human and mouse genomes, but determining their target genes remains challenging. Experimental efforts that examine one locus at a time have established likely causal cCREs for several genes, but they cannot deliver genome-scale analyses. Here we develop Linkreg, a Bayesian model that infers cCRE-gene links by relating the gene expression level to the nearby cCREs' epigenomic states that are derived from a combination of epigenomic features in multiple cell types. In simulations, Linkreg achieves significantly higher power than existing methods, while rigorously controlling false discoveries. On the mouse VISION and human EpiMap datasets, Linkreg identifies many cCRE-gene links with high confidence, which are further validated in external experiments, such as CRISPR screening. In summary, Linkreg connects cCREs to their target genes through effective modeling of epigenomic states across cell types, leading to a more accurate and interpretable characterization of gene regulatory network.

2:00-2:15 PM

Quantification, Association Testing, and Optimal Design for Intratumor Heterogeneity in Multiregion Tumor Studies

Jianxin Shi, National Cancer Institute

Multiregion genomic studies have recently been conducted to characterize intratumor heterogeneity (ITH), to model tumor evolution and to evaluate the association between ITH and clinical variables. To maximize the power of detecting ITH associations in a multiregion study, it is important to appropriately choose the number of patients and the number of tumor samples for these patients within a given budget. For a continuous and a binary variable, we develop statistical methods to find the optimal design for both pre-collection and post-collection scenarios. We show that the optimal design depends on parameters that determine the relative magnitude of the variance of the ITH estimator and the between subject variance of the underlying ITH. In practice, these parameters can be estimated based on pilot studies. The methods are flexible for different types of genomic data. We illustrate the methods using the data from a multiregion study that investigates the ITH of somatic copy number alterations of lung adenocarcinoma.

2:15-2:30 PM

Tracking Pan-Cancer Genomic Evolution Using a Regularized Likelihood Model

Yujie Jiang, The University of Texas MD Anderson Cancer Center

Understanding cancer subclone structure is imperative to providing biological insight in tumor evolution and advancing precision cancer treatment. Recent subclonal reconstruction methods require heavy computing resources, prior knowledge of the number of subclones, and extensive postprocessing. These drawbacks can be addressed by using a regularized likelihood modeling approach, which is novel to the field. Therefore, we develop CliP for fast subclonal reconstruction. We find CliP is robust and accurate in both whole genome (PCAWG, n = 1,993) and whole exome sequencing (TCGA, n = 7,711) data spanning >30 cancer types. The reconstructed tumor subclonality associates with clinical subtypes and prognosis. We also characterize driver mutations in subclonality pan-cancer and find prostate cancer presents the highest frequency of subclonal driver mutations. In summary, our results represent a significant methodological advancement in subclonal reconstruction, and highlight the importance of knowledge in tumor subclone structure. We provide a practical guide and pragmatic pan-cancer resource for the growing interest in studying cancer evolution.

2:30-2:45 PM

Accurate Identification of Locally Aneuploid Cells by Incorporating Cytogenetics Information in Single Cell Data Analysis

Ziyi Li, The University of Texas MD Anderson Cancer Center

Single-cell RNA sequencing is becoming an increasingly common tool to investigate the cellular population and patients? outcomes in cancer research. However, due to the sparse data and the complex tumor microenvironment, it is challenging to identify neoplastic cells that play important roles in tumor growth and disease progression. This challenge is exaggerated in the research of blood cancer patients, from whom the neoplastic cells can be highly similar to normal cells. Here, we present partCNV/partCNVH, a statistical framework for rapid and accurate detection of aneuploid cells with local copy number deletion or amplification. Our method uses an expectation-maximization (EM) algorithm with mixtures of Poisson distributions while incorporating cytogenetics information to guide the classification (partCNV). When applicable, we further improve the accuracy by integrating a hidden Markov model for feature selection (partCNVH). We evaluate the performance of the proposed methods using extensive simulation studies and three scRNA-seq datasets from patients with blood cancers.

2:45-3:00 PM

The Winner's Curse Under Dependence: Challenges and a Comparative Study

Stijn Hawinkel, VIB/Ghent University

Omics experiments yield parameter estimates for thousands of features, of which usually only the top features get reported. Yet these estimates are subject to selection bias or winner's curse: part of the reason why they are extreme is a high estimation error. Proposed solutions include conditional likelihood, bootstrapping and empirical Bayes methods. The most popular member of the latter class is Tweedie's formula, but it has been claimed to be invalid in presence of dependence between the estimates. We demonstrate, however, that the problem lies with inaccurate density estimation under dependence, and restore competitive performance for Tweedie's formula by stabilizing the density estimation through bagging. Next, we benchmark the different selection bias correction methods in a simulation study. Finally, we apply these corrections to the comparison of single- and multigene prediction models for Brassica napus phenotypes from a field trial. It seems as though the best single gene model outperforms the multigene model, and could serve as a predictive marker. Yet correction for selection bias shows that the multigene model is likely still the best performer.

3:00-3:15 PM

Cross Species Integration Pipeline of Omics Data

Ziyue Wang, National Institutes of Health

In past decades, studies have shown the success in using animal models for identifying important loci to understand mechanism of complex diseases. Then, a main question facing the community is ?How do results from your work impact human??. Considering the highly concordant genome shared between mammals and human, integrating genetic data cross species can facilitate the discovery of disease-related genes. In additional, single data modality, such genomics, cannot capture the entire biological complexity of most human diseases. Meanwhile, environmental factors also play an important role in disease etiology. Therefore, statistical methods for leveraging multi-omics data as well as cross species is an emerging need. We proposed a statistical model, for cross species integration on improving functional inference and prioritizing of genomic signals of interests. The methods integrate omics-data across multiple species based on the latent factor analysis and taking advantages of enrichment analysis to establish interpretability. Application to a study in cancer between human and animal populations demonstrates the utility of the approach.

3:15-3:30 PM

Spectral Clustering Using Gene Expression and Histology Identifies Disease-Relevant Spatial Domains in Spatially Resolved Transcriptomics

Kyle Coleman, University of Pennsylvania

Spatially resolved transcriptomics (SRT) provides an unprecedented opportunity to integrate gene expression with histology and spatial location information when studying disease. A prominent goal in SRT data analysis is the identification of spatial domains through clustering of SRT spots. Here we present SpeCTrE, a deep learning-based spectral clustering algorithm for the grouping of SRT spots into spatial domains that are distinct with respect to gene expression and histology. SpeCTrE first employs HIPT to extract spot-level histology features while capturing the long-range histological dependencies among spots. The gene expression, spatial location, and HIPT features are used to construct an adjacency matrix representing the similarity of each pair of spots. Using this adjacency matrix and a multilayer perceptron, SpeCTrE obtains spot-level feature vectors that are used as input for the k-means clustering algorithm. Through analyses of SRT datasets from cancerous tissue sections and extensive benchmark evaluations, we show that SpeCTrE outperforms state-of-the-art spatial clustering methods in separating spots into disease-relevant spatial domains.

36. CONTRIBUTED PAPERS: EXACT METHODS, NETWORK MODELS, COMPUTATIONAL GEOMETRY, DATA VISUALIZATION

Chair: Wenyu Gao, Harvard University

1:45-2:00 PM

Much Ado About Almost Nothing: Statistical Methods for Analyzing Limited Data

Stephen Looney, Augusta University

Even in this era of "big data," applied statisticians are still faced with situations in which there appear to be no useful data, or data of only limited usefulness. For example, when attempting to find a confidence interval for a binomial proportion, the sample may contain no ?successes.? Such a scenario could be encountered when attempting to estimate the incidence of an extremely rare side effect associated with the administration of a newly developed drug. Other statistical inference situations in which the data appear to be of limited (or no) value include estimating an odds ratio when one of the cells in the 2x2 table is empty and performing a correlation analysis when there are observations below the limit of detection of the measuring device(s). In this presentation, the author illustrates each of these scenarios with real data he has encountered in his consulting practice, and describes how valid statistical methods that can be used to perform the desired analysis.

2:00-2:15 PM

Statistical Shape Analysis of Shape Graphs with Applications to Retinal Blood-Vessel Networks

Aditi Basu Bal, Florida State University

This paper provides theoretical and computational developments in statistical shape analysis of shape graphs and demonstrates them using analysis of complex retinal blood-vessel (RBV) network data. The shape graphs are represented by a set of nodes and a set of edges (planar articulated curves) connecting some of these nodes. The goals are to utilize shapes of edges and locations of nodes to: (1) characterize full shapes, (2) quantify shape differences, and (3) model statistical variability. We develop a mathematical representation, elastic Riemannian shape metrics, and associated tools for graph matching, shape geodesics, shape summaries, and shape modeling. One key challenge here is the registration of nodes across large graphs, and we develop a novel multi-scale representation of shape graphs to handle this challenge. We utilize the concepts of effective resistance to cluster nodes and elastic shape averaging of edge curves to change graph details while maintaining overall structures. Registration is then performed by bringing graphs to similar scales before matching. We demonstrate these ideas on Retinal Blood Vessel (RBV) networks from the STARE and DRIVE databases.

2:15-2:30 PM

The Full Picture: Innovative Software Tools for Data Insight Generation

Erya Huang, Bayer AG

Clinical trial databases contain abundant information with complex interrelations, increasing the demand for innovative data visualization methods that provide an insightful overview of study results. In this presentation, we will provide an overview of 7 data visualization tools, all developed for explorative data analyses at the Bayer Biostatistics Innovation Center. These apps enable the user to interactively display information of interest in the form of very powerful static and dynamic graphs. The 7 apps cover all key aspects of a clinical trial, i.e., patient characteristics and data quality, safety (AdEPro, DetectoR, and elaborator), and efficacy; and, more specifically, identification of outcome-relevant subgroups (Subgroup Explorer), identification of associations between variables and developments over time (Megaplots), and structured benefit-risk assessment. New insights generated will help better understand the complex relationships underpinning the clinical data and further drive the understanding of disease development, patient journey, and the impact of the treatments studied on the patient's quality of life.

2:30-2:45 PM

Estimating Gaussian Graphical Models of Multi-Study Data with Multi-Study Factor Analysis

Katherine Shutta, Harvard School of Public Health; Brigham and Women's Hospital and Harvard Medical School

Network models are powerful tools for gaining new insights from complex biological data. Most lines of investigation in biology involve comparing datasets in the setting where the same predictors are measured across multiple studies or conditions (multi-study data). Consequently, the development of statistical tools for network modeling of multi-study data is a highly active area of research. Multi-study factor analysis (MSFA) is a method for estimation of latent variables (factors) in multi-study data. In this work, we generalize MSFA by adding the capacity to estimate Gaussian graphical models (GGMs). Our new tool, MSFA-X, is a framework for latent variable-based graphical modeling of shared and study-specific signals in multi-study data. We demonstrate through simulation that MSFA-X can recover shared and study-specific GGMs and outperforms a graphical lasso benchmark. We apply MSFA-X to analyze maternal response to an oral glucose tolerance test in targeted metabolomic profiles from the Hyperglycemia and Adverse Pregnancy Outcomes (HAPO) Study, identifying network-level differences in glucose metabolism between women with and without gestational diabetes mellitus.

2:45-3:00 PM

Club Exco: Clustering Brain Extreme Communities from Multi-Channel EEG Data

Matheus Guerrero, King Abdullah University of Science and Technology

Current methods for clustering nodes in a brain network are determined by measures computed from the entire range of values of the EEG signals. One limitation of these measures is that they do not distinguish whether the signals are dependent only at large amplitudes or over the entire range of values. We develop the Club Exco method for clustering brain extreme communities to overcome this shortcoming. Club Exco uses spherical k-means applied to the angles derived from extreme amplitudes of EEG signals to cluster EEG data. A cluster center is then considered an extremal prototype, revealing a community of EEG nodes sharing the same extreme behavior: large amplitudes from one node are in tandem with large amplitudes of the others. Non-extreme-value methods cannot identify this important feature. Hence, Club Exco serves as an exploratory tool for classifying EEG channels into mutually asymptotically dependent groups. It provides insights into how the brain network organizes itself during a seizure in contrast to a normal state. Club Exco reveals differences in the organization of the alpha band in the brain network of an epileptic patient compared to coherence-based methods.

3:00-3:15 PM

Multi-Scale Wavelet Coherence with Applications to Brain Connectivity

Haibo Wu, Statistics Program, CEMSE Department, King Abdullah University of Science and Technology

The goal in this paper is to develop a novel statistical approach to characterizing functional interactions between channels in a brain network. Wavelets are effective for capturing transient properties of non-stationary signals. Wavelets give a multi-scale decomposition of signals and thus can be few for studying potential cross-scale interactions between signals. In this paper, we develop scale-specific sub-processes of a multivariate locally stationary wavelet stochastic process. Under this proposed framework, a novel cross-scale dependence measure and its estimation are developed, and it provides a measurement for dependence structure of components at different scales of multivariate time series. The extensive simulation studies are conducted to demonstrate the theoretical properties hold in practice. The proposed cross-scale analysis is applied to the electroencephalogram (EEG) data to study alterations in the functional connectivity structure in children diagnosed with attention deficit hyperactivity disorder (ADHD). Our approach identified novel interesting cross-scale interactions between channels in the brain network.

3:15-3:30 PM

Interpretable AI for Relating Brain Structural and Functional Connectomes

Haoming Yang, Statistical Science Department, Duke University

One of the central problems in neuroscience is understanding how brain structure relates to function. Naively, one can relate direct connections of white matter fiber tracts between brain regions of interest (ROIs) to increased co-activation in the same pair of ROIs, but the link between structural and functional connectomes (SCs and FCs) has proven to be much more complex. To learn a realistic generative model characterizing population variation in SCs, FCs, and the SC-FC coupling, we develop a graph auto-encoder that we refer to as Staf-GATE. We train Staf-GATE using data from the Human Connectome Project (HCP) and show state-of-the-art performance in predicting FCs and joint generation of SCs and FCs. In addition, as a crucial component of the proposed approach, we provide a masking-based algorithm to extract interpretable inference about SC-FC coupling. Our interpretation methods identify important cross-hemisphere and right-hemisphere SC subnetworks for coupling. We demonstrate that coupling subnetworks vary considerably with gender and measures of cognition.

Monday, March 20, 2023 | 3:45-5:30 PM

37. RECENT DEVELOPMENTS IN LONGITUDINAL OMICS DATA ANALYSIS

Organizer/Chair: Gen Li, University of Michigan

3:45-4:10 PM

Strain Genetic Association Studies within the Human Microbiome

Curtis Huttenhower, Harvard T.H. Chan School of Public Health

A rich ecosystem of statistical methods for microbial community epidemiology has evolved that can associate community features with phenotypes, covariates, and exposures across human (and other) populations. At the same time, computational methods for processing metagenomic sequences have improved, yielding increasingly precise features describing community taxa, functions, and strains. Due to the complexity of microbial genomics, however, quantitative methods for genetic epidemiology have not yet been adapted to analyze strain features in these contexts, particularly at scale. I will discuss a suite of statistical models developed to test several different ways in which microbial strain biology can be linked to population phenotypes, including 1) enriched (or depleted) gene variants, 2) nonrandom phylogenetic assortment (using Phylogenetic Generalized Linear Mixed Models, PGLMMs), and 3) strain-specific pathway carriage. These methods have been implemented in the R/Bioconductor package anpan, which has been validated using a variety of synthetic community datasets and applied to identify strain variants consistently associated with colorectal cancer.

4:10-4:35 PM

Estimation of Multi-Block Time Evolving Network Models

George Michailidis, Washington University in St. Louis

In this talk, we discuss multi-block network models wherein the block respects a topological ordering and their dynamics are governed by autoregressive processes. We derive the maximum likelihood estimator for the posited model for Gaussian data in the high-dimensional setting based on appropriate regularization schemes for the parameters of the block components. To optimize the underlying non-convex likelihood function, we develop an iterative algorithm with convergence guarantees. We establish theoretical properties of the maximum likelihood estimates, leveraging the decomposability of the regularizers and a careful analysis of the iterates. The performance of the estimation scheme is evaluated on synthetic data and on epidemiological data from Covid-19.

4:35-5:00 PM

Robust Bayesian Analysis of Longitudinal Omics Data

Cen Wu, Department of Statistics, Kansas State University

The increasing availability of longitudinal omics data has provided unprecedented opportunities to develop novel high dimensional statistical methods that can identify important omics features associated with the disease phenotype. In this talk, we present a general framework of robust Bayesian analysis for omics data with repeated measurements. The proposed Bayesian variable selection method is robust to extreme observations and heteroscedastic model errors while accommodating the structural sparsity in the omics data. The fully Bayesian nature of the developed methods enables inferences through efficient posterior sampling. In simulation, we demonstrate the superiority of the proposed methods in terms of identification accuracy and prediction performance. Real data analysis has shown that the proposed method identifies markers with biologically meaningful implications.

5:00-5:25 PM

Microbial Risk Score for Capturing Microbial Characteristics, Integrating Multi-Omics Data, and Predicting Disease Risk

Huilin Li, New York University

Motivated from the polygenic risk score framework, we propose a microbial risk score (MRS) framework to aggregate the complicated microbial profile into a summarized risk score that can be used to measure and predict disease susceptibility. Specifically, the MRS algorithm involves two steps: 1) identifying a sub-community consisting of the signature microbial taxa associated with disease, and 2) integrating the identified microbial taxa into a continuous score. The first step is carried out using the existing sophisticated microbial association tests and pruning and thresholding method in the discovery samples. The second step constructs a community-based MRS by calculating alpha diversity on the identified sub-community in the validation samples. Moreover, we propose a multi-omics data integration method by jointly modeling the proposed MRS and other risk scores constructed from other omics data in disease prediction. The proposed MRS framework sheds light on the utility of the microbiome data for disease prediction and multi-omics integration, and provides great potential in understanding the microbiome's role in disease diagnosis and prognosis.

38. THE EMERGENCE OF BIOSTATISTICS IN THE MID-20TH CENTURY

Organizer/Chair: Joel Greenhouse, Carnegie Mellon University

3:45-4:10 PM

The International Biometric Society - A History Through the Archives

Lynne Billard, University of Georgia

The International Biometric Society (Society) was founded at Woods Hole on September 6 1947. The formation and structure of the Society through the eyes of the Archival records are outlined. Details of events at Woods Hole as well as key events impacting the Society in later years are described. Brief looks at the regional structure, especially its first region, the Eastern North American Region (ENAR), are described. Publications, beginning with the flagship journal Biometrics and its transfer from the American Statistical Association to the Society, as well as International Biometric Conferences are presented; other factors such as constitutional issues, office management and international affiliations enter into this framework.

4:10-4:35 PM

Biometry at the Early NIH

Christopher Phillips, Carnegie Mellon University

When the first statisticians were recruited to the National Institutes of Health (NIH) in 1947, their work was primarily in biometry, particularly collaborating with cancer researchers with the design and analysis of dose-response and bioassay studies within the laboratories of the National Cancer Institute. Over the 1950s, this group, led by demographer Harold Dorn and statistician Jerome Cornfield, would fundamentally change the scope and meaning of biometry. This paper looks at the role of biometrics in the context of mid-century NIH research, and in particular at how a small group of statistically-minded individuals—few of whom had any formal prior training in biometry, vital statistics, or public health methods—could fundamentally change how we understand the place of statistical inference and measurement in modern medicine.

4:35-5:00 PM

The Role of Biostatistics in the Development of the Field of Regulatory Science at FDA

Robert O'Neill, Retired from FDA

The growth of the biostatistical field in the pharmaceutical area was fueled by a combination of public health crises that occurred over the years, the recognition that the science of statistics and applied statistics was critical for implementing the regulatory standards for clinical evidence of efficacy and safety of new drugs, especially the design, analysis and interpretation of clinical trials, and the surveillance of product safety. Most important was the FDA’s willingness to provide human resources to support and build a program of professional statisticians that were strong in methodology, consulting, negotiating, and affecting drug regulatory decisions. In this talk I will review some of the regulatory issues that arose in the 1960s and 1970s and stimulated this growth, and the events that triggered statistical contributions, the people who played a key role in addressing those issues, and how this work impacted the development of the practice of biostatistics in the remainder of the twentieth and twenty first century.

5:00-5:25 PM

DISCUSSANT

Janet Wittes, Emerita Statistics Collaborative, Inc.

39. DATA INTEGRATION METHODOLOGIES AND ALGORITHMS: SOME NEW DEVELOPMENTS

Organizer: Peisong Han, University of Michigan
Chair: Xu Shi, University of Michigan

3:45-4:10 PM

Doubly Robust Estimators for Generalizing Treatment Effects on Survival Outcomes from Randomized Controlled Trials to a Target Population

Shu Yang, North Carolina State University

In the presence of heterogeneity between the randomized controlled trial (RCT) participants and the target population, evaluating the treatment effect solely based on the RCT often leads to biased quantification of the real-world treatment effect. To address the problem of lack of generalizability for the treatment effect estimated by the RCT sample, we leverage observational studies with large samples that are representative of the target population. This paper concerns evaluating treatment effects on survival outcomes for a target population. We propose a semiparametric estimator through the guidance of the efficient influence function. The proposed estimator is doubly robust because it is consistent for the target population estimands if either the survival model or the weighting model is correctly specified and is locally efficient when both are correct. We also employ the nonparametric method of sieves for flexible and robust estimation of the nuisance functions and show that the resulting estimator retains the root-n consistency and efficiency. Empirical studies confirm the theoretical properties of the proposed estimator and show it outperforms competitors.

4:10-4:35 PM

Distributed Multi-Site Latent Class Analysis (dMLCA) with Application to Subphenotyping of Multisystem Inflammatory Syndrome in Children

Yong Chen, University of Pennsylvania

Latent class analysis (LCA) is an unsupervised learning method that can be used to disentangle complex conditions by identifying their unobserved subphenotypes. Performing LCA on data from multiple clinical sites enables more generalizable and reliable findings compared to a single-site study. In this talk, we describe a Distributed Multi-site Latent Class Analysis (dMLCA) algorithm, which allows the learning of subphenotypes from binary, categorical, and/or count data without sharing patient-level data across sites while accounting for the between-site heterogeneity. Our dMLCA produces subject-specific posterior probabilities of subphenotypes that account for between-site heterogeneity and subject-level covariates when they are available. These inferred subject-specific posterior probabilities of subphenotypes can be directly factored into considerations of clinical decision-making. Furthermore, we applied this novel method to conduct subphenotyping of the multisystem inflammatory syndrome in children (MIS-C) using data from nine academic medical centers including 864 MIS-C patients extracted from 3,549,894 pediatric patients from March 2020 to December 2021.

4:35-5:00 PM

Information Sharing for Efficient Inference from Different Data Sources

Emily Hector, North Carolina State University

A fundamental aspect of statistics is the integration of data from different sources. Classically, Fisher and others were focused on how to integrate homogeneous sets of data. More recently, the question of if data sets from different sources should be integrated is becoming more relevant. The current literature treats this as a yes/no question: integrate or don't. Here we take a different approach, motivated by information-sharing principles coming from the shrinkage estimation literature. In particular, we deviate from the binary, yes/no perspective and propose a dial parameter that controls the extent to which two data sources are integrated. How far this dial parameter should be turned is shown to depend on the informativeness of the different data sources as measured by Fisher information. This more-nuanced data integration framework leads to relatively simple parameter estimates and valid tests/confidence intervals. We demonstrate both theoretically and empirically that setting the dial parameter according to our recommendation leads to more efficient estimation compared to other binary data integration schemes. This work is joint with Ryan Martin.

5:00-5:25 PM

Improved Prediction Mean Squared Error for Linear Regression by Integrating External Model Results

Peisong Han, University of Michigan

When integrating external study summary information into an internal study of interest to improve model fitting, it is inevitable that estimation bias is introduced due to study population heterogeneity and/or the uncertainty associated with the external information that may be very difficult to account for. Such estimation bias may diminish or even undo the benefit of integrating external information because of increased mean squared error (MSE). For linear regression models, we develop a data integration procedure based on James-Stein shrinkage that guarantees to improve the prediction MSE after information integration, regardless of the population or sample size heterogeneity across studies.

40. RECENT ADVANCES IN OPTIMAL SAMPLING DESIGN AND ESTIMATION USING VALIDATED AND ERROR-PRONE DATA

Organizer: Noorie Hyun, Kaiser Permanente Washington Health Research Institute
Chair: Yates Coley, Kaiser Permanente Washington Health Research Institute

3:45-4:10 PM

Optimal Multi-Wave Validation Study in a Multi-National HIV Research Cohort

Gustavo Amorim, Vanderbilt University Medical Center

Kaposi's sarcoma (KS) is a leading type of cancer affecting people living with HIV (PLWH). Patients with a low CD4 count at diagnosis are, in particular, at greater risk of developing and dying from KS. Since the use of antiretrovirals (ART) has decreased the incidence of KS over time, we hypothesize that it has further decreased within the Treat-All era. To evaluate this question, we used Electronic Medical Records (EMR) from nearly 300,000 patients collected from sites within the IDEA East Africa and CCASANet networks, from 2010 to 2019. However, as EMRs are error-prone, we designed a multi-wave validation study to validate critical variables in the EMR dataset using patients' medical charts. Our validation design targets patients that are more informative about the research question and works by minimizing the variance of design-based estimators related to the parameters of interest. It works in two waves, allowing the design to be updated for the second wave using data from patients selected in the first wave. We validated key records from 500 patients in each wave and incorporated the chart review into the EMR dataset to estimate the incidence of KS among PLWH

4:10-4:35 PM

Design and Analysis Strategies with "Secondary" Use Data

Ran Tao, Vanderbilt University Medical Center

The growing availability of observational databases like electronic health records (EHR) provides unprecedented opportunities for secondary use of such data in biomedical research. However, these data can be error-prone and need to be validated before use. It is usually unrealistic to validate the whole database due to resource constraints. A cost-effective alternative is to implement a two-phase design that validates a subset of patient records that are enriched for information about the research question of interest. In this talk, I will discuss proper statistical approaches to analyze such two-phase studies, which can efficiently use the information in the unvalidated data in Phase I and address the potential biased validation sample selection in Phase II. I will demonstrate the advantages of the proposed methods over existing ones through extensive simulations and an application to an ongoing HIV observational study.

4:35-5:00 PM

Optimal Validation-Sample Design: Reconciling Model-Based and Design-Based Solutions?

Thomas Lumley, Department of Statistics, University of Auckland

For a linear regression model under two-phase sampling, optimal designs when the model is assumed correct involves sampling only from extreme values of predictor or outcome, but when the model is not assumed correct, the optimal design involves sampling the entire range of predictor and outcome. Previous work and the local asymptotic minimax theorem suggest that for some form of model misspecification the transition between these two optima must happen over a family of contiguous alternatives to the true model. We explore what this transition looks like.

5:00-5:25 PM

An Augmented Likelihood Approach that Incorporates Error-Prone Auxiliary Data into a Survival Analysis

Noorie Hyun, Kaiser Permanente Washington Health Research Institute

In this big data era, we can easily observe substantial amounts of clinical data in large observational studies or electronic health records (EHR). Data accuracy can vary according to measurement methods. For example, self-reported medical history can include bias, such as recall bias or response bias. In contrast, gold standard diagnostic tests are less likely to be biased but may not be available on all individuals in a large prospective study due to cost or participant burden. We are motivated to study what benefit we can gain by augmenting analyses of the gold standard disease outcome with error-prone self-reported disease diagnoses in regression for time-to-disease onset. The proposed model addresses left-truncation and interval-censoring in time-to-disease onset outcomes while correcting errors in self-reported disease diagnosis in a joint likelihood for the gold standard and error-prone outcomes. The proposed model is applied to the Hispanic Community Health Study/ Study of Latino data to quantify risk factors associated with diabetes onset.

41. KICKSTART YOUR CAREER: HOW TO MAXIMIZE THOSE EARLY YEARS

Organizer: Xingruo Zhang, University of Chicago
Chair: Tianyi Sun, Vanderbilt University

3:45-4:10 PM

Kickstart your Career: How to Maximize Those Early Years

Panelists:
Xiangrong Kong, Johns Hopkins University
Torri Simon, Boston Scientific
Andrew Spieker, Vanderbilt University
Briana Stephenson, Harvard T.H. Chan School of Public Health
Lucy D'Agostino McGowan, Wake Forest University

The transition from graduate school to navigate the unknowns of the job market is challenging for every new statistician. Proper training, efficient networking, and building a professional profile are some of the early initiatives to prepare graduate students for this change. Statisticians are trained in modeling and data analysis; however, the real-world job market requires skills beyond technical knowledge, including communication, presentation, leadership, and collaborative skills, as well as pitching one's ideas and goals, and being able to advocate for oneself. As a group of emerging statisticians, CENS would like to fill this gap and invite early-career statisticians to discuss the unique challenges that early-career statisticians might face in a new work environment and how to deal with them.

Our panel includes both academic and industry statisticians at the MS and PhD level, several of whom graduated in the last five years. Their valuable insights and mentoring guidance will be useful for newly-emerging statisticians to build a path to kickstart their careers achieving their goals.

42. ESTIMAND AND SENSITIVITY ANALYSIS IN CLINICAL TRIALS: IMPLEMENTING THE ICH-E9 (R1) GUIDELINES

Organizer: Lu Mao, University of Wisconsin-Madison
Chair: Tuo Wang, University of Wisconsin-Madison

3:45-4:10 PM

Estimands: The New Bedrock of Drug Development

Frank Bretz, Novartis

With the publication of ICH E9(R1) Addendum there has been an uptake of the estimand framework and principled missing data approaches. Clinical questions of interest and associated estimands are now specified at the initial stages of planning a clinical trial, leading to a better alignment of trial objectives, trial design, data collection and method of analysis. In this presentation we will share lessons learned and best practices on the implementation of the estimand framework in pharmaceutical development, based on the experiences through an internal cross-functional and cross-divisional working group that encompasses various estimand initiatives as well as external initiatives and engagements.

4:10-4:35 PM

General Pairwise Comparisons: Toward Estimand-Driven Analysis

Lu Mao, University of Wisconsin-Madison

The win ratio, win odds, and net benefit are all effect-size measures calculated by comparing every patient in the treatment to every one in the control. Such generalized pairwise comparisons (GPC) allow flexible ranking of the outcomes according to their clinical severity, and are increasingly used in the analysis of prioritized composite endpoints consisting of death and nonfatal events. Most GPC-based methods, however, are first and foremost concerned with construction of estimator, oftentimes at the cost of a dubious estimand that either depends on the censoring distribution or lacks causal interpretation. This needs to change in the wake of the recently released ICH-E9(R1) Addendum, which places estimand construction at the core of the design and analysis of clinical trials. In this talk, I summarize two broad-based approaches to defining meaningful estimands that naturally entail GPC-like estimators --- a nonparametric one by restricting the time frame and a semiparametric one by modeling the time-dependent comparison result.

4:35-5:00 PM

Causal Inference for Comprehensive Cohort Studies

Daniel Scharfstein, University of Utah School of Medicine

In a comprehensive cohort study of two competing treatments (A,B), clinically eligible individuals are first asked to enroll in a randomized trial and, if they refuse, are then asked to enroll in a parallel observational study in which they can choose treatment according to their own preference. We consider estimation of comprehensive cohort causal effect (i.e., the difference in mean potential outcomes had all patients in the comprehensive cohort received treatment A vs. treatment B). We illustrate our methodology using data from the BARI (Bypass Angioplasty Revascularization Investigation) randomized trial and observational registry to evaluate the effect of percutaneous transluminal coronary balloon angioplasty (PTCA) versus coronary artery bypass grafting (CABG) on 5-year mortality.

5:00-5:25 PM

DISCUSSANT

Michael Rosenblum, Johns Hopkins Bloomberg School of Public Health

43. CONTRIBUTED PAPERS: GRAPHICAL MODELS/METHODS

Chair: Rupam Bhattacharyya, University of Michigan

3:45-4:00 PM

Microbial Network Analyses: Hubs, Multiple Conditions, and Complex Design

Shilan Li, Georgetown University

Microbiome network analyses are widely used to reveal interacting patterns among microbial taxa. Moreover, a hub, a taxonomy connected with many other taxa, may be informative to understand the complex interacting pattern. Further, many studies include multiple conditions, offering an opportunity to increase power for identifying shared edges while identifying condition-specific edges. Following Liang et al. (2014), we develop a computationally efficient and statistically powerful method for network analysis with hub detection in multiple-condition studies. Extensive simulations showed that our approach well-controlled false positive rates and greatly improved the power of detecting hubs and shared edges. Existing methods based on Gaussian graphic models are computationally prohibitive and often detect too many spurious edges by choosing tuning parameters using BIC. We further extended our approach to survey samples with a complex design. We show that ignoring sampling weights inflated type I error rate and reduced the power. We demonstrate our approach by analyzing the oral microbiome data with three lung cancer subtypes and subcohort samples from the PLCO study.

4:00-4:15 PM

Two-Sample Bayesian Causal Directed Acyclic Graphs for Observational Zero-Inflated Count Data

Junsouk Choi, Department of Statistics at Texas A&M University

Observational zero-inflated count data arise in a wide range of areas such as economics and biology. A common research question in these areas is to identify causal relationships by learning the structure of a sparse directed acyclic graph (DAG). While structure learning of DAGs has been an active research area, existing methods do not adequately account for excessive zeros and hence are not suitable for modeling zero-inflated count data. Moreover, in many scientific settings, it is often interesting to study differences in causal networks for data collected from two experimental groups (control vs treatment). To explicitly account for zero-inflation and identify differential causal networks, we propose a novel Bayesian differential zero-inflated negative binomial DAG (DAG0) model. Our main theorem proves that the causal structure of DAG0 is fully identifiable from purely observational data, using a general proof technique applicable beyond our DAG0. Bayesian inference based on parallel-tempered Markov chain Monte Carlo is developed to efficiently explore the multi-modal posterior landscape. We show the utility of DAG0 through extensive simulations and real data analysis.

4:15-4:30 PM

Heterogeneous Network Analysis of Disease Outcomes via Mining Electronic Medical Record Data

Jiping Wang, Yale University

High-quality electronic medical records have advanced the development of human disease networks (HDNs), especially from the clinical outcome perspective. Based on clinical outcome HDNs, healthcare practitioners can carry out more effective and efficient practices with improved clinical outcomes for patients. Clinical treatment outcomes (e.g., length of stay) are usually count data that deviate from normal. Therefore, the Gaussian graphical model (GGM) fails and novel statistical models are required. We apply Poisson Log-normal-based GGM and regression-based approaches to accommodate the non-normal count data and clinical covariates that can reveal sparser and more fundamental HDNs. We extend the analytical frameworks that allow heterogeneous populations and can automatically determine the number of groups by including a fusion penalty. Effective EM-based algorithms are developed and show satisfactory performance via extensive simulations and comparisons. The heterogeneity HDN analysis of the Medicare claims data based on clinical outcomes not only demonstrates the practical applicability of the proposed approach but also has higher clinical practical value.

4:30-4:45 PM

Probabilistic Graphical Modeling under Heterogeneity

Liying Chen, University of Michigan, Ann Arbor

Probabilistic graphical models such as Gaussian graphical models (GGM) are widely used to visualize and interpret complex dependencies in multi-variate and high-dimensional biomedical datasets. Most current probabilistic GGM-based methods assume homogeneous samples which limits the applicability of these models to model heterogeneity across samples that is routinely present in many scientific contexts. We propose a flexible approach called Graphical Regression (GraphR) which allows for covariate-dependent graphs and thus enables incorporation of metrics of sample heterogeneity. Our regression-based method provides a functional mapping from the covariate space to precision matrix for different types of heterogeneous graphical model settings. GraphR is flexible in incorporating different scales of covariates; imposes sparsity in both edge and covariate selection and computationally efficient via use of variational Bayes algorithms. We explore the comparative efficacy of our method thorough various simulation settings and demonstrate the versatility of the method through applications to diverse multi-omic and spatial transcriptomics datasets to study regulatory networks.

4:45-5:00 PM

Targeted Gene Expression Inference with Heterogeneous Gaussian Graphical Models

Qiong Wu, University of Pennsylvania

The gaussian graphical model (GGM) is a powerful tool for studying gene expression networks of complex diseases. Transfer learning benefits the estimation of GGMs in a target dataset by leveraging useful information from related source studies. However, patients with the same complex disease can behave differently with distinct latent genetic patterns, which can result in heterogeneity across patients from different datasets. With such heterogeneity, an informative source study may be identified as irrelevant in prior or lead to negative transfer if improperly incorporated. Hence, we developed a multi-class transfer learning framework to facilitate the learning of target GGMs by integrating multiple source datasets under population heterogeneity. Within each latent class, certain similarities between source and target GGMs are shared to allow for knowledge transfer. We provided the estimation procedure when subgroup structures are known or need to be determined data-dependently. We conducted extensive simulations and applications in breast cancer patients to demonstrate the superior performance of the proposed framework compared to classic single-site learning.

5:00-5:15 PM

Selecting a Significance Level in Sequential Testing Procedures for Community Detection

Ian Barnett, University of Pennsylvania

While there have been numerous sequential algorithms developed to estimate community structure in networks, there is little available guidance and study of what significance level or stopping parameter to use in these sequential testing procedures. Most algorithms rely on prespecifiying the number of communities or use an arbitrary stopping rule. We provide a principled approach to selecting a nominal significance level for sequential community detection procedures by controlling the tolerance ratio, defined as the ratio of underfitting and overfitting probability of estimating the number of clusters in fitting a network. We introduce an algorithm for specifying this significance level from a user-specified tolerance ratio, and demonstrate its utility with a sequential modularity maximization approach in a stochastic block model framework. We evaluate the performance of the proposed algorithm through extensive simulations and demonstrate its utility in controlling the tolerance ratio in single-cell RNA sequencing clustering by cell type and by clustering a congressional voting network.

5:15-5:30 PM

Interactive Network Clustering and Investigation of Complex Multivariate Association Matrices with Association Subgraphs

Yaomin Xu, Vanderbilt University Medical Center

Making sense of networked multivariate association patterns is vitally important to many areas of high-dimensional analysis. Unfortunately, as the data-space dimensions grow, the number of association pairs increases in O(n2); this means traditional visualizations such as heatmaps quickly become too complicated to parse effectively. Here we present associationSubgraphs: a new interactive network clustering and visualization method to quickly and intuitively explore high-dimensional association datasets using network percolation and clustering. The goal is to provide efficient investigation of association subgraphs, each containing a subset of variables with stronger and more frequent associations among themselves than the remaining variables outside the subset, by showing the entire clustering dynamics and provide subgraphs under all possible cutoff values at once. Particularly, we apply associationSubgraphs to a phenome-wide multimorbidity association matrix generated from an electronic health record (EHR) and provide an online, interactive demonstration for exploring multimorbidity subgraphs.

44. CONTRIBUTED PAPERS: BAYESIAN SPATIAL/TEMPORAL MODELING

Chair: Tianjian Zhou, Colorado State University

3:45-4:00 PM

A Time-Dependent Poisson-Gamma Model for Recruitment Forecasting in Multicenter Studies

Armando Turchetta, McGill University

Forecasting recruitments is a key component of the monitoring phase of multicenter studies. Yet, deterministic models mainly based on trial investigators' recruitment assumptions are still used. A Bayesian approach built on a doubly stochastic Poisson process, known as the Poisson-Gamma model, was introduced to address the lack of a strong and consistent statistical methodology in this field. This approach is based on the modeling of enrollments as a Poisson process where the recruitment rates are assumed to be constant over time and to follow a common Gamma prior distribution. However, the constant-rate assumption is a restrictive limitation that is rarely appropriate for applications in real studies. In this presentation, we illustrate a flexible generalization of this methodology which allows the enrollment rates to vary over time by modeling them through B-splines. We show the suitability of this approach for a wide range of recruitment behaviors in a simulation study and by estimating the recruitment progression of the Canadian Co-infection Cohort (CCC).

4:00-4:15 PM

A Bayesian Zero-Inflated Beta-Binomial Model for Longitudinal Timeline Followback Data

Chun-Che Wen, Medical University of South Carolina

Timeline followback (TLFB) is often used in addiction research to monitor recent substance use. Typically, TLFB data comprise binomial counts that exhibit overdispersion and zero inflation. Motivated by a 12-week study evaluating a new smoking cessation treatment, we propose a Bayesian zero-inflated beta-binomial model for longitudinal TLFB data. The model comprises a mixture of a point mass that accounts for zero inflation and a beta-binomial distribution for the number of days abstinent in the past week. Because treatment effects appear to wane during the study, we introduce random changepoints for each treatment group to reflect group-specific behavioral changes over time. The model also includes fixed and random effects that capture group- and subject-level slopes before and after the changepoints. We can accurately predict the weekly probability of abstinence and test whether the treatment groups experience changepoints at the same time. We demonstrate that the proposed model outperforms alternative models, including the ordinary beta-binomial. In our application, we show that the new treatment has a short-term positive effect that tapers off after week 9.

4:15-4:30 PM

A Bayesian Approach for Modeling Variance of Intensive Longitudinal Biomarker Data as a Predictor of Health Outcomes

Mingyan Yu, University of Michigan, Ann Arbor

The development of intensive longitudinal biomarker data has led to the development of methods to predict health outcomes and facilitate precision medicine. Intensive biomarker data is measured at a high frequency and typically results in several hundred to several hundred thousand observations per individual measured over minutes, hours, or days. Often In longitudinal studies, the primary focus is on the means of trajectories, and the variances are treated as nuisance parameters, although they may also be informative for the outcomes. We propose a Bayesian hierarchical model to jointly and simultaneously model the outcome and the predictors. To model the variability of predictors and deal with the high intensity of data, we develop subject-level penalized splines, which allow sharing of information across individuals for both the residual variability and the penalized random effects. Then different levels of variability are extracted and incorporated into the outcome models to make an inference. We demonstrate an application of the joint model using bio-monitor data including hertz-level heart rate data from a study on social stress.

4:30-4:45 PM

Bayesian Kernel Machine Regression for Count Data: Modeling the Association Between Social Vulnerability and COVID-19 Deaths in South Carolina

Fedelis Mutiso, Medical University of South Carolina

Motivated by a study examining associations between county-level social vulnerability and COVID-19 deaths, we develop a Bayesian kernel machine regression model for count data. The model uses a kernel machine representation to examine interactive and nonlinear associations between 15 social vulnerability variables and COVID-19 death rates. The method produces county-specific vulnerability effects, makes predictions for future county exposure profiles, and quantifies the relative importance of each social vulnerability variable. To capture spatiotemporal heterogeneity, we introduce spatial effects, county-level covariates, and smooth functions of time modeled via cubic B-splines. Restricted spatial regression is used to investigate spatial confounding. For Bayesian computation, we propose an efficient Pólya-Gamma data augmentation algorithm that relies on easily sampled Gibbs steps. We conduct a simulation study and apply the method to a study of COVID-19 deaths in South Carolina. Results indicate that social vulnerability had the greatest impact on deaths in the northwest portion of the state and was driven primarily by higher rates of poverty and lower education levels.

4:45-5:00 PM

Inferring HIV Transmission Patterns from Viral Deep-Sequence Data via Latent Spatial Poisson Processes

Fan Bu, University of California, Los Angeles

Viral deep-sequencing data play a crucial role in understanding disease transmission flows by providing in-depth evidence. To fully utilize these data and account for uncertainty in phylogenetic analysis, we propose a spatial Poisson process model to uncover population-level HIV transmission flows. We represent pairs of individuals with viral sequences as typed points, with coordinates representing covariates and point types representing unobserved transmission statuses (link and direction), which are informed by scores obtained through phylogenetic analysis that summarizes viral sequence data evidence. Our method jointly infers pairwise latent transmission statuses and the transmission flow surface on the source-recipient space. Unlike existing methods, our framework avoids heuristic pre-classification of pairwise transmission statuses, instead learning them probabilistically through a Bayesian inference scheme, meanwhile enjoying significant computational speed-up thanks to a continuous spatial process design. In a case study on viral sequence data from Uganda, our method captures high-resolution age structures in HIV transmission and brings valuable insights.

5:00-5:15 PM

WITHDRAWN Bayesian Estimation of Spatial GLM with Pólya-Gamma Data Augmentation Algorithm

Xuan Ma, Case Western Reserve University

Count data over geographic regions are reported in various fields of research, including the study of climate, biology, economics, etc. Sampling from the posterior distribution for modeling count data with logistic regression in the Bayesian approach is inefficient due to the lack of conjugate prior. The data augmentation algorithm has been widely used since Albert and Chib (1993) proposed truncated normal as augmented variables for Bayesian probit regression. The analogous development for Bayesian logistic regression by Polson, Scott and Windle (2013) introduced the Pólya-Gamma data augmentation algorithm. In this study, we exploit the Pólya-Gamma scheme and propose an efficient Bayesian MCMC sampling approach for spatial binomial regression and spatial negative binomial regression. Spatial effects for areal and point-level models are considered with the CAR model and Squared Exponential model as examples. Simulation studies highlight the good performance of the algorithm with the accuracy of the estimation and efficiency of the sampling method compared to other approaches.

45. CONTRIBUTED PAPERS: EDUCATION, CONSULTING, AND HEALTH POLICY

Chair: Michele Guindani, UCLA

3:45-4:00 PM

A Pattern Discovery Algorithm for Pharmacovigilance Signal Detection

Anran Liu, University at Buffalo

Safety of medical products continues to be a major public health concern worldwide. Spontaneous Reporting Systems (SRS), such as the FDA Adverse Event Reporting System (FAERS), are critical tools in the post-marketing evaluation of medical product safety. A variety of approaches have been developed for identification of adverse events using data that reside in FAERS and other SRS databases. We propose a pattern discovery algorithm, named Modified Detecting Deviating Cells (MDDC) algorithm, for the identification of adverse events when the database is represented as an I x J contingency table. The MDDC procedure is based on the standardized Pearson residuals of the pairs of potential adverse event-drug combination, allowing the change of scale from categorical to interval/ratio scale. The method is 1) easy to compute; 2) considers the relationship between the different adverse events; 3) depends on a data driven cutoff. We discuss various methods for cut-off identification and study its performance via simulation. We apply the method on a statin drug class data set downloaded from FAERS.

4:00-4:15 PM

The Role of the Modern Statistician in the Era of Easily Disseminated Statistical Knowledge

Liam O'Brien, Colby College

The Dunning-Kruger effect describes the phenomenon whereby those with limited knowledge in a particular domain overestimate their ability within that domain. The foundation of the scientific method requires hypothesis testing, and often with statistical analysis. The demand for statistical expertise and statistical training has outstripped the number of statisticians available. Burgeoning fields such as data science are less likely to involve statisticians presuming their training is strong enough. Some may even question the value of statistical training. The role of statistics should be more important than ever. However, as the lines among fields of expertise become blurred, data-driven recommendations are often being disseminated by practitioners that may not be qualified. We discuss (1) whether all indexed journals should have statisticians on staff to review all papers (2) whether all grant organizations should require statisticians on staff to review all grants and (3) whether undergraduate and graduate statistical coursework should emphasize the importance of collaboration and consultation. Finally, we will discuss whether statistics has a marketing problem.

4:15-4:30 PM

Team Science as a Guide for Methodological Decision-Making to Quantify Trends in the Drug Overdose Case Fatality Rate

Emily Slade, University of Kentucky

Changes in the drug overdose case fatality rate (the proportion of drug overdoses resulting in death) over time could signal that drug overdose response resources need to be increased or redistributed in communities. Quantifying changes in the overdose case fatality rate is a methodological challenge since the numerator and denominator are both random variables measured over time. The choice of Bayesian beta regression as the methodological framework as well as the discovery of previously-unknown seasonal trends in the overdose case fatality rate would not have been possible without several features of the collaborative study team including: (1) team formation and role definition as a foundation of trust, (2) discussion of competing and common goals among team members, (3) cyclical communication between domain experts and methodological experts through key phases, and (4) integration of multiple biostatisticians on the project. This project highlights the role that collaborative biostatisticians can play to weave team science principles through a methodologically-challenging biomedical problem with the ultimate goal of leveraging results for community and policy-based action.

4:30-4:45 PM

A Swarm of Lines in the Life of a Simple Linear Regression Model

Marepalli Rao, University of Cincinnati

A simple linear regression model has three parameters: the intercept; slope; error variance. A sample (Yi, Xi), I = 1, 2, ? , n of size n is drawn to estimate the parameters of the model. The least squares estimators of the slope and intercept are BLUE. Given any two data points (Yi, Xi) and (Yj, Xj) with distinct X?s, we have a line joining them. We will then have a swarm of such lines, which are inferior to the least squares lines. We will combine some of these lines to get the least squares line. We will also introduce the concept of a pair ((Yi, Xi), in which Yi is influential and Xi is a leverage point combining both the notions.

4:45-5:00 PM

Using Simulations to Model PSA Screening Policies

Holly Hartman, Case Western Reserve University

There are major racial disparities in prostate cancer (PCa) where, relative to non-Hispanic white (NHW) men, non-Hispanic Black (NHB) men are more likely to be initially diagnosed with distant metastases, be diagnosed at a younger age, and have higher prostate cancer specific mortality (PCSM). Early-stage prostate cancer has a relatively high survival rate, yet more advanced prostate cancer has a much worse survival rate, making prevention of metastatic disease a critical component to preventing PCSM. While prostate specific antigen (PSA) testing is highly sensitive in detecting PCa, including early-stage disease, there are also significant concerns of overdiagnosis, the possibility of diagnosing indolent tumors, and overtreatment. Here, we explore using agent-based modeling to examine the impacts of changing prostate cancer screening recommendations on the future PCSM rates by race. We use publicly available data to set parameters for the simulations including screening rates, disease progression rates, and mortality rates. Due to the earlier onset of PCa in NHB men, age-based recommendations may not be appropriate and may widen disparities.

5:00-5:15 PM

Quantifying Racial Disparities in Kidney Graft Failure Rates Using US Registry Data with Federated Learning Algorithms

Dazheng Zhang, University of Pennsylvania Perelman School of Medicine

In United States, differential kidney transplant access to minority groups is found to be associated with worse health outcomes. For example, black patients may have worse graft failure rate to white patients. It was hypothesized that some of the differences between the graft failure rate of white and black patients is attributable to differential access to hospital cares. We quantified racial disparities in kidney graft failure rates that are associated with different access to transplant centers. Specifically, we developed a counterfactual model to assess whether the graft failure rates differ if the black patients had been admitted to the same distribution as the white patients were admitted. To enable the analysis using multi-site data without sharing patient-level data, we proposed a communication-efficient federated algorithm for a flexible time-to-event model. Our algorithm only requires the aggregated data to be shared across centers in three rounds of communications. Extensive simulation studies were conducted to evaluate the performance of the proposed algorithm. The federated algorithm was applied to the data from 149 kidney transplant centers.

5:15-5:30 PM

Secure Federated Hospital Profiling Using Real-World Data

Jiayi Tong, University of Pennsylvania

Hospital profiling allows for quantitative comparisons of healthcare providers' quality of care. Providing high-quality details on patient health conditions, EHR data hold a great promise for hospital profiling. We proposed a secure federated learning framework, called dGEM (decentralized Generalized mixed Effects Model), which tackles the challenges in multisite EHR-based hospital profiling: patient-level data sharing dilemma and case-mix situation. We validated the framework with a centralized register data from 187 centers. The framework achieved nearly identical results as the gold standard method. We then investigated the variation in hospital performances measured by COVID mortality for two pandemic periods with a federated EHR data of 191,682 patients from 12 international sites in the OHDSI network. We found that 8 sites improved the standardized mortality rate in the second period. The proposed federated framework is highly applicable and potentially transformative for health outcome research.

46. CONTRIBUTED PAPERS: FUNCTIONAL DATA HYPOTHESIS TESTING AND ESTIMATION

Chair: Yuyan Wang, NYU Grossman School of Medicine

3:45-4:00 PM

Shape-Constrained Estimation in Functional Regression with Bernstein Polynomials

Rahul Ghosal, University of South Carolina

Shape restrictions on functional regression coefficients such as non-negativity, monotonicity, convexity or concavity are often available in the form of a prior knowledge or required to maintain a structural consistency in functional regression models. A new estimation method is developed in shape-constrained functional regression models such as scalar-on-function regression (SOFR), function-on-scalar regression (FOSR), and function-on-function regression (FOFR) using Bernstein polynomials. Theoretical results establish the asymptotic consistency of the constrained estimators under standard regularity conditions. A projection based approach provides point-wise asymptotic confidence intervals for the constrained estimators. Numerical analysis using simulations illustrate improvement in efficiency of the estimators from the use of the proposed method under shape constraints. Two applications include i) modeling a drug effect in a mental health study via shape-restricted FOSR and ii) modeling subject-specific quantile functions of accelerometry-estimated physical activity in the Baltimore Longitudinal Study of Aging (BLSA).

4:00-4:15 PM

Regression and Alignment for Functional Data and Network Topology

Danni Tu, University of Pennsylvania

The human functional brain network dynamically reorganizes during adolescence. Changes in mesoscale topology can be assessed by modularity and participation coefficient, two diagnostics which capture the community structure of the brain network. By proportionally thresholding the network edges, we obtain a sequence of diagnostics for each threshold, resulting in diagnostic curves that describe network structure at multiple scales. Previous methods that evaluate network diagnostic curves have relied on permutation-based or pointwise comparisons, which are less powerful and less informative than comparisons of curves in their entirety. We propose a functional regression framework that addresses biases introduced by systematic differences in the distribution of edge strengths between networks, which we conceptualize as phase variation in diagnostic curves. Our novel method therefore simultaneously performs regression and curve alignment through an iterative, penalized estimation procedure. The illustrated procedure is widely applicable to domains of neuroscience where the goal is to study heterogeneity among a mixture of function- and scalar-valued measures.

4:15-4:30 PM

Equivalence Test for Two Mean Curves of Functional Data

Haiou Li, Georgetown University

The paper considers hypothesis testing for comparing mean functions under the function data setting. Existing methods rely on dimension reduction approaches through decompositions and/or truncations, including spectral decompositions of covariance functions in ANOVA and functional principal extractions. The dimension reduction eases the complexity of the infinite process but also suffers some aspects of information loss. Meanwhile, when detecting the difference of mean functions, there is no criteria or distance measure to identify the magnitude between two functions that might suggest dissimilarity in underlying (e.g., biological) settings. Thus, we propose an equivalence test with a given size of difference for comparing two mean functions based on the distance of random elements in Hilbert space. The asymptotic properties of the testing statistic are also obtained. The size of the difference is chosen to incorporate both theoretical and practical significance.? ? The extensive simulation studies show the superior performance of the proposed method, and the real data analyses using the proposed statistic are valid and powerful and offer new insight into the practical problem.

4:30-4:45 PM

Two Sample Test for Eigendecompositions of Functional Data

Angel Garcia de la Garza, Albert Einstein College of Medicine

We develop statistical procedures to compare the eigendecomposition from two samples of functional data. We first introduce an appropriate test when the observations on both groups are independent and extend it to the case of paired functions. Our procedure tests the covariance matrix of FPCA scores rather than comparing the eigendecomposition directly. Our work focuses on an experiment in which a trained mouse reached for a food pellet after an auditory cue. We are motivated to understand whether activation patterns in the motor cortex remain constant as the mouse repeatedly performs the reaching movement. Our results suggest trial-to-trial variation in the latent activation patterns that can't be attributed to sampling noise. Thus it is important to account for trial-to-trial variability when deriving activation patterns from neural-spike data.

4:45-5:00 PM

Projection-Based Two-Sample Inference for Multivariate Sparse Functional Data

Salil Koner, Duke University

Modern longitudinal studies collect multiple outcomes as the primary endpoints to understand the complex dynamics of the diseases. Oftentimes, especially in clinical trials, the joint variations among the multidimensional responses play a significant role in assessing the differential characteristics between two or more groups, rather than drawing inferences based on a single outcome. Enclosing the longitudinal design under the umbrella of sparse functional data, we develop a projection-based two-sample test to identify the difference between the typical multivariate profiles. The test is built upon functional principal component analysis to reduce the dimension of the functions while preserving the correlation between them. Finite-sample numerical studies demonstrate that the test maintains type-I error, and is powerful to detect significant group differences, compared to the state-of-the-art procedures. The test is carried out on the TOMORROW study of individuals at high risk of mild cognitive impairment due to Alzheimer's disease to detect differences in the cognitive test scores between the pioglitazone and the placebo groups.

5:00-5:15 PM

Covariance Estimation for Mixed Type Functional Data Using Semiparametric Gaussian Copula

Debangan Dey, National Institute of Mental Health

Digital diaries available through smartphones enable real-time real-life tracking of mood, energy, and many other self-assessed states of homeostatic systems. These data can be treated as multivariate functional observations of mixed type (e.g. continuous, binary, ordinal, and truncated. For example, individual&rsquo;s mood is typically reported on a Likert scale of 1-7 (ordinal) 4 times a day over a week. We define Generalized Latent Non-paranormal Process to develop a semiparametric Gaussian Copula-based approach for covariance estimation of mixed functional data. Semiparametric Gaussian Copula mechanism assumes that observed process is generated by i) monotonically transforming marginals of latent multivariate Gaussian process and ii) dichotimizing/truncating the marginals. In addition, our method also predicts latent functional principal component (FPC) scores. We demonstrate this method to model daily/weekly covariance structure of mood (ordinal) reported 4 times a day over two weeks from 592 participants of National Institute of Mental Health family study and investigate how latent mood FPC scores vary across healthy controls and participants with mood disorders.

5:15-5:30 PM

mfMR: Multivariable Functional Mendelian Randomization Incorporating Longitudinal Data

Hanfei Xu, Boston University School of Public Health

Mendelian randomization is a useful approach to investigate causal relationships between exposures and complex traits in epidemiological studies that can potentially overcome confounding. However, the MR method still lacks a way of incorporating multiple time-varying exposures into analyses. We propose models that incorporate a functional data analysis approach to handle multiple time-varying exposures under a multivariable MR framework. We also introduce the concept of mean functional exposure, yielding interpretable causal effect estimates. Our simulation study demonstrates that the proposed models perform better than alternative methods utilizing only a single measurement, in terms of both statistical power and bias of the effect estimate. We apply our models with data from the Framingham Heart Study Offspring cohort to study the respective direct causal effects of body mass index and waist-hip ratio on various bone health related measures. Our method advances the research of causal inference by making better use of longitudinal information from multiple exposures, and thus can provide more insights into the relationship between exposures and the outcome of interest.

47. CONTRIBUTED PAPERS: NEUROIMAGING ANALYSIS

Chair: Eunchan Bae, University of Pennsylvania

3:45-4:00 PM

Scalable Bayesian Image-on-Scalar Regression for Population-Scale Neuroimaging Data Analysis

Yuliang Xu, University of Michigan

Image-on-scalar regression (ISR) has been widely used to study the associations between clinical outcomes and brain imaging data. Bayesian ISR has the flexibility of modeling sparsity and spatial dependence through various prior specifications and can provide straightforward uncertainty quantification on the model parameters in posterior inference. However, Bayesian ISR is very challenging when analyzing large-scale imaging data due to the limited scalability of standard posterior computation methods. In this work, we adopt Gaussian process priors on the regression parameters and introduce salience region indicators in Bayesian ISR. We develop a scalable posterior computation algorithm based on stochastic gradient Langevin dynamics and memory mapping techniques. The proposed method can directly make spatial posterior inferences on brain activation regions when the brain masks are inconsistent across individuals, a common problem with fMRI data. We demonstrate the advantages of the proposed method via extensive simulations and analysis of the UK Biobank task fMRI data.

4:00-4:15 PM

Bayesian Longitudinal Tensor Response Regression for Modeling Neuroplasticity

Alec Reinhardt, Department of Biostatistics, UT MD Anderson Cancer Center

A major interest in longitudinal neuroimaging studies involves investigating voxel-level neuroplasticity changes due to clinical factors. We propose a novel Bayesian tensor response regression approach for longitudinal imaging data, which pools information across voxels in order to infer significant neuroplasticity changes while adjusting for covariate effects. The proposed method employs low-rank decomposition to reduce dimensionality and preserve spatial configurations of voxels during estimation, and allows for feature selection via a joint credible regions approach. Advantages of the proposed approach in terms of prediction and feature selection compared to routinely used voxel-wise regression are highlighted via extensive simulation studies. Subsequently, we apply the method to an Aphasia dataset to examine group and individual level neuroplasticity measured from task functional MRI data. Our analysis revealed spatial and temporal differences in neuroplasticity across treatment interventions and age. By contrast, standard voxel-wise regression failed to detect any significant neuroplasticity changes after multiplicity adjustments.

4:15-4:30 PM

Mediation Analysis for High-Dimensional Mediators and Outcomes with an Application to Multimodal Imaging Data

Zhiwei Zhao, University of Maryland, College Park

Multimodal neuroimaging data have attracted increasing attention for brain research. Integrated analysis of multimodal neuroimaging data and behavioral or clinical measurements provides a promising approach for investigating the underlying neural mechanisms of different phenotypes. However, such an integrated data analysis is intrinsically challenging due to the complex interactive relationships between the multimodal multivariate imaging variables. We propose a multivariate-mediator and multivariate-outcome-mediation model (M6) to extract the latent mediation patterns and estimate the causal mediation effects based on a bi-cluster graph approach. We develop a computationally efficient algorithm for bi- cluster structure estimation and inference to identify the mediation patterns. We validate M6 with an extensive simulation and compare it with conventional methods. We further apply the proposed method to multimodal imaging data from the Human Connectome Project to investigate the effect of systolic blood pressure on whole-brain imaging measures for the regional homogeneity of the blood oxygenation level-dependent signal through the cerebral blood flow.

4:30-4:45 PM

Sparse Partial Logistic Tensor Regression with Application to Neuroimaging Data

Dayu Sun, Emory University

Tensor data, i.e., multi-dimensional arrays, have been increasingly prevalent in biomedical studies, e.g., in neuroimaging applications. The complications of tensor data analysis include the high-dimensionality and intrinsic structures within tensor data. The regression of a continuous response on tensor predictors has been well studied in recent years, but there is relatively limited literature on the tensor regression of a binary outcome, whose methods may suffer from a lack of uniqueness in estimation. In this work, we propose a Sparse Partial Logistic Tensor Regression method for modeling binary outcomes on both tensor and vector/scalar predictors. We utilize mode-wise penalized manifold optimization to achieve dimension reduction and sparsity in tensor coefficient estimation that may improve the prediction performance. Extensive simulation studies show our proposal achieve satisfactory performance under various scenarios and outperforms existing methods. We apply our proposal to study the association between the diagnosis of posttraumatic stress disorder and brain connectivity matrices derived from functional magnetic resonance imaging data from a mental health study.

4:45-5:00 PM

Hypothesis Tests of Spatial Enrichment in Brain-Behavior Association Studies

Sarah Weinstein, University of Pennsylvania

Neuroimaging-based studies of neurodevelopment, neurological disorders, and mental health typically involve collecting a combination of different brain measurements (e.g., structure and function) as well as behavioral measures. In such studies, we are often interested in quantifying and spatially mapping brain-behavior associations, and ultimately, evaluating whether these associations are especially strong--or spatially ?enriched?--within sub-regions of the brain (e.g., functional networks). In this work, we propose and compare multivariate test statistics to quantify brain-behavior associations, including burden and sequential kernel association test statistics. We find that these multivariate statistics produce more interpretable maps of brain-behavior associations compared with more commonly used univariate approaches. To examine spatial enrichment, we propose an adaptation of Gene Set Enrichment Analysis (GSEA, Subramanian et al., 2005), which involves permutation-based testing to control type I error levels. We conduct data-driven simulation studies and analyses using neuroimaging data from a longitudinal study of neurodevelopment.

5:00-5:15 PM

DeepCombat: A Statistically-Motivated, Hyperparameter-Robust, Deep Learning Approach to Harmonization of Neuroimaging Data

Fengling Hu, University of Pennsylvania

MRI images and subsequent image-derived radiomic features, including cortical thicknesses, exhibit pronounced technical artifacts due to differences in MRI scanner magnet strength or manufacturer. These scanner effects can obscure biological effects of interest or decrease the reproducibility of findings, especially as analysis methods become more powerful. Image harmonization methods seeking to remove scanner effects are important for mitigating these issues. We present DeepCombat, a deep-learning harmonization method based on a conditional variational autoencoder architecture and the ComBat harmonization method. DeepCombat uses this structure to learn and remove individual-level scanner effects in both feature means and residuals while accounting for the relationships between features. We apply this method to a large cognitive-aging cohort and find that DeepCombat outperforms existing methods in removing scanner effects from cortical thickness measurements while preserving biological heterogeneity.

48. CONTRIBUTED PAPERS: JOINT MODELS FOR LONGITUDINAL AND SURVIVAL DATA

Chair: Chia-Rui (Jerry) Chang, Harvard T.H. Chan School of Public Health

3:45-4:00 PM

Joint Model Framework for Data with a Terminal Event, Recurrent Event, and Associated Longitudinal Measure: A Bayesian Approach

Emily Damone, University of North Carolina at Chapel Hill

Many methods exist to jointly model either recurrent and related terminal survival events or longitudinal measures and related terminal survival event. However, few methods exist which can account for the dependency between all three outcomes of interest. We propose a joint model which uses subject-specific random effects to connect the survival (terminal and recurrent time-to-event) data with a longitudinal outcome measure. Proportional hazards models with shared frailties are used to analyze recurrent and terminal events while a separate (but dependent) set of random effects are utilized in a generalized linear mixed model to analyze longitudinal measures. Random effects are related through a multivariate normal distribution with structured covariance matrix. The proposed joint modeling approach provides a flexible model that can be utilized in a wide range of health applications. We evaluate the model through simulation studies as well as through an application to patients from the Atherosclerosis Risk in Communities (ARIC) study.

4:00-4:15 PM

Incorporating Cross-Sectional Information into a Joint Model of Longitudinal and Survival Data Using a Power Prior

Juned Siddique, Northwestern University Feinberg School of Medicine

Joint modeling of longitudinal data and survival outcomes is a useful approach for understanding how cardiovascular risk factor trajectories over time affect the development of cardiovascular disease (CVD) later in life. Prospective cardiovascular cohort studies provide an appropriate source of information to address these questions, but a limitation is that many of these cohort studies are relatively small and/or have a low number of events. Conversely, there exist large representative cross-sectional surveys that provide an abundant source of information on the relationship between cardiovascular risk factors and CVD, but do not contain information on risk factor trajectories. In this talk, we describe a flexible Bayesian approach for obtaining more precise inferences by incorporating cross-sectional risk factor data and its association with outcomes into a joint model through the use of a power prior. We use longitudinal data from the Coronary Artery Risk Development in Young Adults (CARDIA) cohort study and cross-sectional data from the Third National Health and Nutrition Examination Survey (NHANES) Linked Mortality File.

4:15-4:30 PM

A Multi-State Model for Smoking Cessation and the Risk of Lung Cancer in the Alpha-Tocopherol, Beta-Carotene Cancer Prevent Trial

Sung Duk Kim, National Cancer Institute

Cancer Epidemiologists are often interested in characterizing the relationship between smoking behavior and cancer in large populations. Using data from the Alpha-Tocopherol, Beta-Carotene Cancer Prevent (ATBC) we consider the challenge of relating longitudinal smoking behavior to the risk of lung cancer in a population of Finnish males smoking at the time of randomization. We develop a discrete-time multi-state model of smoking behavior and lung cancer incidence through specification of a state-space incorporating both components. We consider the relevant smoking history as including whether they smoked or not over 3-month intervals and a latent ?quit? state. We introduce random effects for each of the transition probabilities to incorporate heterogeneity across individuals. Likelihood and Bayesian approaches are proposed for estimation. The model is applied to examine the complex relationships between patterns of smoking behavior (including smoking status, average weekly cigarette consumption, and quit status) on lung cancer risk.

4:30-4:45 PM

Joint Model of Longitudinal Data and Time-to-Event Outcome with Bayesian Approach: Simulation Study and Application to Women's Health

Chen Chen, University of Toronto

Joint models have been used to combine information from longitudinal predictors and time-to-event outcomes. However, these models often restrict to using information about individual-level mean trajectories of a longitudinal biomarker, e.g., random intercepts and slopes to predict hazard functions for time-to-event data (e.g., Cox proportional hazard models). Here we extend this approach to consider whether the individual-level variability of the longitudinal biomarker contains information about the hazard in addition to the mean trends. In a Bayesian framework, we use a Gaussian mixture model to estimate random intercept and slopes for the subject-level mean, and a separated mixture model for the log of the subject-level variance. These are then included as predictors in an accelerated failure time model for a survival outcome with a log link function. The proposed model is coded in Rstan and evaluated with simulations to compare with two-stage competing methods. We then apply the proposed model to predict mean and variance of C-reactive protein as longitudinal predictor and time to diabetes as the outcome in the Study of Women?s Health Across the Nation (SWAN).

4:45-5:00 PM

Dynamic Prediction Frameworks for Generalized Joint Modeling of Multivariate Functional Mixed Model and Time to Event Data

Dongrak Choi, Duke University

Parkinson’s disease (PD) is a complicated neurodegenerative disease impairing multiple domains. Many clinical studies on PD collect multivariate longitudinal outcomes using the MDS-Unified Parkinson's Disease Rating Scale (MDS-UPDRS) to fully explore the impairment caused by the disease. The MDS-UPDRS, most commonly used in PD, measures longitudinal ordinal outcomes using a Likert scale. To accommodate these ordinal outcomes, we propose a generalized joint model based on a multivariate functional mixed model that simultaneously models the multiple longitudinal ordinal outcomes and time to patients movement disabilities. We also develop a Bayesian approach for parameter estimation and a dynamic prediction framework for predicting target patients future risk of a survival event based on their current outcome. The proposed model is evaluated by simulation studies and applied to the PPMI, an observational study assessing the progression of clinical features of PD.

5:00-5:15 PM

Dynamic Prediction Using Backward Joint Model with Multivariate Nonlinear Longitudinal Biomarker Trajectories

Yi Yao, The University of Texas MD Anderson Cancer Center

In follow-up data of patients undergoing renal transplantation, the longitudinal trajectories of prognostic biomarkers of kidney function are likely to be nonlinear, especially when the measurement time approaches the time of the terminal event. Existing dynamic prediction methods are still challenged by the statistical and computational difficulties of jointly modeling multiple nonlinear longitudinal profiles and time-to-event in face of right censoring. We tackle these challenges using the backward joint modeling approach. We fit the model by first estimating the conditional distribution of time-to-event given other baseline covariates using a piecewise exponential survival model, then a penalized spline linear mixed effect model is used to describe the evolution of each biomarker conditioning on the smoothed bivariate surface of terminal event time and measurement time, with potential between-marker correlation specified in the random effects. The procedure is implemented via penalized pseudo-likelihood and the Expectation-Maximization algorithm.

5:15-5:30 PM

Multistate Processes for Alcohol Use Disorder and Related Health Issues: An Application to Kentucky Medicaid Database

Yuchen Han, University of Louisville

Continuous time multi-state models are theoretically ideal to handle transitions between clinical states. However computational burden and preprocessing steps have so far limited its use of investigating behavioral health issues in big data. Our study for alcohol use disorder and mental health is the first attempt to establish a robust computational workflow on such data (i.e., KY Medicaid database from 2012-2019). We start out from raw ICD9 and ICD10 codes for all related disorders to form a smoothed version by filtering out unexpected transition patterns on patient-level longitudinal records. After model-fitting, we interpret in detail the estimated log transition rates and visually identify some interesting patient trajectories obtained from fitted transition probabilities. Next, we propose a few further refinements through a Bayesian nonparametric framework which aims to find latent subgroups based on patient trajectories and hence address sources of variation potentially unexplained by common covariates.

Tuesday, March 21, 2023 | 8:30-10:15 AM

49. RECENT ADVANCES IN INFERENTIAL STATISTICAL METHODS IN GENOMICS

Organizer: Saptarshi Chakraborty, State University of New York at Buffalo
Chair: Ronglai Shen, Memorial Sloan-Kettering Cancer Center

8:30-8:55 AM

SUITOR: Selecting the Number of Mutational Signatures Through Cross-Validation

Bin Zhu, Division of Cancer Epidemiology and Genetics, National Cancer Institute

For de novo mutational signature analysis, the critical first step is to decide how many signatures should be expected in a cancer genomics study. An incorrect number could mislead downstream analyses. Here we present SUITOR (Selecting the nUmber of mutatIonal signaTures thrOugh cRoss-validation), an unsupervised cross-validation method that requires little assumptions and no numerical approximations to select the optimal number of signatures without overfitting the data. In vitro studies and in silico simulations demonstrated that SUITOR can correctly identify signatures, some of which were missed by other widely used methods. Applied to 2,540 whole-genome sequenced tumors across 22 cancer types, SUITOR selected signatures with the smallest prediction errors and almost all signatures of breast cancer selected by SUITOR were validated in an independent breast cancer study. SUITOR is a powerful tool to select the optimal number of mutational signatures, facilitating downstream analyses with etiological or therapeutic importance.

8:55-9:20 AM

Using the “Hidden Genome” to Mine Mutation Contexts Across the Cancer Genome to Map Tumor Site of Origin

Saptarshi Chakraborty, State University of New York at Buffalo

It is increasingly common clinically for cancer specimens to be examined using techniques that identify somatic mutations. In principle, these mutational profiles can be used to characterize tumor cancer type. However, most observed mutations are rare, and indeed in new hitherto unobserved mutations are routinely encountered. To create a viable diagnostic tool we need to harness the information content in this hidden genome of variants for which no direct information is available. To accomplish this we propose a Bayesian multilevel meta-feature regression model to extract the critical information from rare variants in the training data in a way that permits us to also extract diagnostic information from any previously unobserved variants in the new tumor sample. A scalable implementation of the model is obtained by combining a high-dimensional feature screening approach with a group-lasso penalized maximum likelihood approach. We apply the method to three publicly available datasets -- PCAWG whole genome, TCGA whole exome, and MSK-IMPACT targeted cancer genome sequencing. Results show that our approach can harness substantial diagnostic information from the hidden genome.

9:20-9:45 AM

Predicting Cancer Risk from Germline Next-generation Sequencing Data Using a Novel Context-based Variant Aggregation Approach

Zoe Guan, Memorial Sloan Kettering Cancer Center

Most cancer types have a substantial heritable component. However, known risk variants explain only a limited proportion of the estimated heritability of cancer. It has been hypothesized that much missing heritability lies in rare variants not captured by SNP arrays. Rare variants need to be aggregated to achieve sufficient statistical power for detecting associations. We propose a novel context-based variant aggregation approach for extracting signals from rare variants detected through germline whole-exome and whole-genome sequencing. Many studies have shown that the genomic, nucleotide, and epigenetic contexts of somatic variants in tumors are informative of cancer etiology and site of origin. Recently, evidence has also emerged that the contexts of germline variants are associated with cancer risk. Using germline whole-exome data from the UK Biobank, we investigate the predictive value of meta-features aggregating rare variants based on their genomic, nucleotide, and epigenetic contexts for distinguishing cancer cases from controls. We compare the performance of risk models based on known risk variants and risk models that additionally include the meta-features.

9:45-10:10 AM

SpaceX: Gene Co-Expression Network Estimation for Spatial Transcriptomics

Satwik Acharyya, University of Michigan

The analysis of spatially-resolved transcriptome enables the understanding of the spatial interactions between the cellular environment and transcriptional regulation. In particular, the characterization of the gene-gene co-expression at distinct spatial locations in the tissue enables delineation of spatial co-regulatory patterns. To enhance the ability and potential of spatial transcriptomics technologies to drive biological discovery, we develop SpaceX (spatially dependent gene co-expression network), a Bayesian methodology to identify both shared and cluster-specific co-expression network across genes in a spatially structured tissue consisting of different clusters in the form of cell classes or tissue domains. SpaceX uses an over-dispersed spatial Poisson model coupled with a high-dimensional factor model which is based on a dimension reduction technique for computational efficiency. We show via simulations, accuracy gains in co-expression network estimation and structure by accounting for increasing spatial correlation. We have discussed in-depth analysis of two spatial transcriptomics datasets in mouse hypothalamus and human breast cancer.

50. DEEP LEARNING FOR SURVIVAL ANALYSIS

Organizer: Ruiwen Zhou, Washington University in St. Louis
Chair: Lei Liu, Washington University in St. Louis

8:30-8:55 AM

Neural Network on Interval Censored Data with Application to the Prediction of Alzheimer's Disease

Ying Ding, University of Pittsburgh

Alzheimer's disease (AD) is a progressive and polygenic disorder that affects millions of individuals each year. Given that there have been few effective treatments yet for AD, it is highly desirable to develop an accurate model to predict the full disease progression profile based on an individual's genetic characteristics for early prevention and clinical management.

8:55-9:20 AM

Deep Learning with Time-to-event Outcomes

Jon Steingrimsson, Brown University

Deep learning is a class of algorithms that uses multiple layers to create a prediction model. The layers involve an unknown weight vector that is estimated by minimizing a loss function. We extend the deep learning algorithms to handle censoring by replacing the loss function used in the absence of censoring by censoring unbiased loss functions. We discuss properties of these loss functions and practical issues related to implementation of the deep learning algorithms. The performance of the resulting algorithms is evaluated through simulation studies and by analyzing data on cancer patients.

9:20-9:45 AM

Interpreting Neural Networks through Hypothesis Testing

Francesca Mandel, University of Pennsylvania

Neural networks are excellent predictive models, but the difficulties of interpretation limit their utility in many fields. Various methods have been proposed to provide insight into the nature of relationships between predictors and outcomes, generally through feature importance rankings and sensitivity analyses. Despite these contributions, statistical inference and formal hypothesis testing of feature associations remain largely unexplored. We propose several approaches to testing based on partial derivatives of the network outputs with respect to specific inputs. We develop tests for assessing association and nonlinearity that can be flexibly applied to a variety of network architectures. These tests enhance the explanatory power of neural networks, which combined with powerful predictive capability, extend the applicability of these models.

9:45-10:10 AM

Neural Network for Clustered Survival Outcomes Using Frailty Model

Ruiwen Zhou, Washington University in St. Louis

A great deal of literature has been established for the analysis of clustered survival experiments, where the subjects within a subgroup of the population share some unobserved effects. A common analysis approach for such data is the frailty model. The shared frailty model under proportional hazard assumption has been widely applied for the analysis of clustered survival outcomes. However, the prediction performance of this method can be poor when the risk function is highly nonlinear. To deal with this issue, we proposed a neural network shared frailty Cox model that replaces the linear risk function in shared frailty Cox model with the output of a feed-forward neural network. The estimation is based on a quasi-likelihood with the use of Laplace approximation. A simulation study suggests that the proposed method works well for practical situations. Furthermore, the method is applied to a set of real data.

51. INCOMPLETE DATA IN NON-TRADITIONAL SETTINGS: ANGLES, FUNCTIONS, AND SHAPES

Organizer: Gregory Matthews, Loyola University Chicago
Chair: Karthik Bharath, School of Mathematical Sciences, University of Nottingham

8:30-8:55 AM

Completion of Partially Observed Curves Using Hot Deck Type Imputation

Gregory Matthews, Loyola University Chicago

Statistical shape analysis of curves is well-developed when curves are fully observed. This work considers partially observed curves and develops methods for curve completion or imputation by leveraging tools from the statistical analysis of shape of fully observed curves, which enables sensible curve completions. On a dataset containing partially observed bovid teeth arising from a biological anthropology application, the method is implemented and classification of the completed teeth is carried out based on a shape distance on the set of curves.

8:55-9:20 AM

Multiple Imputation with Angular Covariates

Benjamin Stockton, University of Connecticut

In this talk, I will discuss analysis of incomplete data with angular components. Angular data arise in various contexts and provide unique challenges for statistical analysis. Due to the geometry of the circle, methods that assume data lie in the standard Euclidean spaces are unable to appropriately analyze angular data. Missing data methods for angular data are under-examined in the literature with few papers addressing the issue directly and fewer solely focused on the missing data problem. Multiple imputation (MI) has been extensively developed to address missing data in a wide variety of contexts including real-valued data and functional data. We propose to fill this gap with a projected normal (PN) imputation method for use within MI. For comparison, we discuss imputation on the angle directly and on the cosin and sin components with standard imputation methods. As a real-world example, at the county-level we analyze adults age-adjusted asthma prevalence in the US based on pollution levels and various meteorological factors including wind direction. The results of the analysis with PN imputations are compared to results from standard imputation methods.

9:20-9:45 AM

Approaches for Extending and Evaluating Multiple Imputation for Functional Data

Adam Ciarleglio, Department of Biostatistics and Bioinformatics, George Washington University

Missing data are a common problem in biomedical research. Valid approaches for dealing with missing data have been proposed and are regularly implemented in applications where the data are exclusively scalar-valued. However, with advances in technology and data storage, biomedical studies are collecting functional data with increasing regularity – and these functional data may be subject to missingness. We propose extensions of multiple imputation with predictive mean matching and imputation by local residual draws as two approaches for handling missing functional data. These methods are compared via a simulation study and applied to data from a study of subjects with major depressive disorder for which both clinical (scalar) and imaging (functional) data are available.

9:45-10:10 AM

DISCUSSANT

Ofer Harel, Department of Statistics, University of Connecticut

52. DECOMPOSING ADMIXED GENOMICS DATA: CELL-TYPE-AWARE ANALYSIS METHODOLOGY ADVANCES

Organizer/Chair: Hao Feng, Case Western Reserve University

8:30-8:55 AM

Decoding Spatially Resolved Tumor Microenvironments with Digital Cytometry

Aaron Newman, Stanford University

In cancer, complex ecosystems of interacting cell types form powerful signaling networks that shape tumorigenesis. A comprehensive understanding of tumor cell states, their co-association patterns, and their impact on clinical outcomes could facilitate new opportunities for disease management and therapeutic intervention. We recently introduced several new computational methods for determining cell states and multicellular ecosystems from bulk, single-cell, and spatially resolved gene expression data. With these methods for “digital cytometry,” cell states and multicellular communities can be profiled at high resolution and massive scale, recovered in expression datasets independent of platform, related to therapy response, and tracked across space and developmental time. In this talk, I will describe these approaches and illustrate their potential to enable exciting new discoveries that are not obtainable from bulk, single-cell, or spatial expression platforms alone.

8:55-9:20 AM

Statistical Methods for Cellular Deconvolution of Human Brain RNA Sequencing Data

Stephanie Hicks, Johns Hopkins Bloomberg School of Public Health

Statistical methods referred to as "cellular deconvolution" have been developed to estimate the relative fractions of cell types in bulk RNA-seq datasets. However, these approaches require reference expression profiles from the underlying cell types, which can be difficult to generate from human postmortem brain tissue. While many approaches have been proposed, the majority produce similar composition estimates. Many of existing reference datasets - regardless of the algorithm employed - are largely non-comparable and produce incorrect estimates of cellular composition. Current algorithms estimate the relative fraction of RNA attributable to each cell type, and not the relative fraction of cell types. To address this, we developed a deconvolution algorithm to capture unbiased estimates of cellular composition in human postmortem RNA-seq data. We demonstrate our approach with both simulated and real bulk RNA-seq data to better determine the relative role of cell type-specific expression in the human brain and their subsequent dysregulation in debilitating brain disorders.

9:20-9:45 AM

Characterizing Clonal Expansion and Microenvironment at Single Cell Resolution

Wenyi Wang, University of Texas MD Anderson Cancer Center

It is now increasingly recognized that both the genetic (evolutionary) and transcriptional (ecological) differences between tumor cells and the interactions between tumor and immune cells play a critical role in metastasis and resistance to systemic therapy. Available statistical models and tools so far have only superficially explored the relationship between various molecules, e.g., DNAs and RNAs that cohabit in cancer cells. Biologically, however, the varying relationship between molecules is a centerpiece of the constantly changing tumor microenvironment. In this talk, I will introduce our integrative model that measures the varying DNA/RNA relationship called TmS, and then further introduce two new model developments DeMix.SC and CliP.SC, which advances our recent methods DeMixT and CliP by incorporating signals in matched single-cell RNA sequencing data from the same patient samples. Our integrative models aim to fill in the gaps between understanding cancer evolutionary dynamics at single cell resolution and at-scale association with clinical outcomes, such as tumor relapse or metastasis.

9:45-10:10 AM

Cell-Type-Aware Statistical Methods for Spatial Transcriptomics

Rafael Irizarry, Dana-Farber Cancer Institute

Spatial transcriptomics technologies permit location-aware high-throughput measurement of gene expression at the single-cell level. A limitation of this technology is that individual measurements may contain contributions from multiple cells, hindering the discovery of cell-type-specific spatial patterns of localization and expression and differential expression analysis. In this talk, I will introduce single-cell RNA-Seq data aided by several informative plots, then describe the statistical challenge in spatial transcriptomics and statistical solutions: RCTD and C-SIDE. Our solution leverages cell-type profiles learned from single-cell RNA-seq and propose a model that permits estimating mixtures and accounting for batch effects.

53. NEW ADVANCES IN MULTIMODAL BRAIN IMAGING INTEGRATION

Organizer: Aiying Zhang, Columbia University/New York State Psychiatric Institute
Chair: Seonjoo Lee, Columbia University and New York State Psychiatric Institute

8:30-8:55 AM

Multimodal Data Fusion Approaches to Neuroimaging Data

Vince Calhoun, GSU/GATech/Emory

Brain imaging technology provides a powerful tool to visualize both functional and structural information. However, most studies ignore the very interesting and complex relationships between the two. Most existing approaches take a one-way approach, rather than evaluating joint inter-modal relationships. Data-driven approaches are particularly informative for studying links between structure and function. We proposed a family of multivariate data-driven approaches, with a focus on independent component analysis and more recently deep learning, which leverage higher order statistics to discover and link together network-level patterns of macroscopic structural and functional MRI data. This allows us to identify links which do not necessarily correspond spatially (e.g., structural changes in one region related to functional changes in other regions). They also provide a network level; perspective on the data, enabling us to identify sets of brain regions that covary together and to evaluate within and between network relationships. We present a variety of examples, including several showing the potential of such approaches to inform us about mental illnesses such as schizophrenia.

8:55-9:20 AM

Interpretable Multimodal Deep Learning for Brain Imaging and Genomics Data Fusion

Wenxing Hu, Tulane University

Deep network-based data fusion models have been developed to capture complex associations between multi-modal datasets such as brain imaging and genomics, resulting in improved diagnosis of mental diseases. However, deep learning models are often difficult to interpret, bringing about challenges for uncovering biological mechanisms using these models. In this work, we develop an interpretable multimodal fusion model to perform automated diagnosis and result interpretation simultaneously. We name it Grad-CAM guided convolutional collaborative learning (gCAM-CCL), which is achieved by combining intermediate feature maps with gradient-based weights. The gCAM-CCL model can generate interpretable activation maps to quantify pixel-level contributions of the input features. Moreover, the estimated activation maps are class-specific, which can therefore facilitate the identification of biomarkers underlying different groups. Finally, we apply and validate the gCAM-CCL model on a brain imaging-genomics study, and demonstrate its applications to both the classification of cognitive function groups and the discovery of underlying biological mechanisms.

9:20-9:45 AM

A High-Dimensional Multi-Exposure Mediation Model to Unravel Brain Structure-Functional Interactions

Aiying Zhang, Columbia University/New York State Psychiatric Institute

Myelin, the dielectric sheath surrounding neuronal axons, maintains the integrity of neural fibers and enhances the speed of propagation of action potentials. It is an essential component for efficient brain functioning, which facilitates long-range neuronal communication processes supporting higher-order cognitive, sensory, and motor functions. We used the ratio maps of the T1-weighted and T2-weighted images as the measurement of myelin and the rs-fMRI for the functional connectivity estimation to understand how structure constrains and shapes function. To estimate the relationship from brain structure, brain function and the behavior, we propose a multi-exposure multivariate mediation analysis for Gaussian distributed data, which allow dependency within and across exposures and mediators. The proposed method has the benefit of potentially: (i) providing more interpretable pathways (i.e., linear combinations of brain regions rather than individual region) and (ii) reducing the number of mediators for estimation.

9:45-10:10 AM

DISCUSSANT

Seonjoo Lee, Columbia University

54. CHALLENGES AND OPPORTUNITIES IN BIOMARKER-DRIVEN TRIAL DESIGN: ADAPTIVE RANDOMIZATION

Organizer: Yeonhee Park, University of Wisconsin-Madison
Chair: Yanhong Zhou, Eli Lilly and Company

8:30-8:55 AM

Adaptive Designs for Precision Medicine: A Review, New Challenges and Innovative Designs

Feifang Hu, George Washington University

Precision medicine proposes the customization of health-care tailored to a patient based on the individual characteristics. Adaptive designs provide effective ways to optimize patients' treatments by incorporating individual information. First, we provide a brief overview of popular adaptive designs that incorporate covariates. Then a general and unified mathematical framework is proposed for adaptive randomization procedures. Currently no design has been proposed in literature to incorporate discrete and continuous prognostic covariates, and predictive covariates simultaneously. To ï¬ll the gap, we proposes a new class of covariate-adjusted response-adaptive procedures to optimize the treatments based on individual covariates under the general framework. The proposed procedure can balance both discrete and continuous covariates to make valid and credible comparisons among treatments. Theoretical properties for the new designs are investigated in some simpliï¬ed scenarios. Extensive numerical studies have been conducted, which demonstrate the superiorities of new procedure over existing designs.

8:55-9:20 AM

Adaptive Randomization for Master Protocols in Precision Medicine

Liwen Wu, Takeda Pharmaceuticals

In the era of precision medicine, especially in oncology and hematology, there have been explosions in knowledge of the disease molecular profiles. New generation of clinical trials have emerged to target patient population within any given tumor type based on specific underlying molecular and biologic characteristics. Among them, umbrella, basket, and platform trials constitute a new generation of clinical trial design defined as master protocol, which allows studying multiple drugs and/or disease indications within a single trial.

These innovative approaches to clinical drug development leads to rapidly revolutionized methodologies, including adaptive randomization, to conduct trials in the biomarker and targeted therapy settings, where the traditional paradigm seems less efficient, lacks cost effectiveness and maybe ethically challenging. In this presentation, we will discuss some recent advancements and new methodologies for adaptive randomization in master protocols, with simulation studies and an illustrative example with a proof-of-concept umbrella study. Overall, these methods provide a more flexible and robust approach in the design and analysis of clinical trials.

9:20-9:45 AM

Using Simulation to Explore Trade-offs In Different Innovative Approaches for Clinical Trials

J. Kyle Wathen, Cytel

When designing a clinical study there are often several approaches for analysis and decision making. It is not always clear which approach can provide the highest likelihood of success or even address the questions the trial is intended to address. In this talk I will discuss the topic of simulation guided design and how to leverage it to help guide the selection of various options for adaptive randomization in clinical trials with biomarker and/or covariates.

9:45-10:10 AM

Personalized Risk-based Screening Design for Comparative Two-Arm Group Sequential Clinical Trials

Yeonhee Park, University of Wisconsin-Madison

In an era of personalized medicine, it is critical to incorporate the patients' characteristics and improve the clinical benefit for patients. The patients' characteristics are incorporated in adaptive randomization to identify patients who are expected to get more benefit from the treatment and optimize the treatment allocation. However, it is challenging to control potential selection bias from using observed efficacy data and the effect of prognostic covariates in adaptive randomization. We propose a personalized risk-based screening design using Bayesian covariate-adjusted response-adaptive randomization that compares the experimental screening method to a standard screening method based on indicators of having a disease. Personalized risk-based allocation probability is built for adaptive randomization, and Bayesian adaptive decision rules are calibrated to preserve error rates. A simulation study shows that the proposed design controls error rates and yields a much smaller number of failures and a larger number of patients allocated to a better intervention compared to existing randomized controlled trial designs.

55. CONTRIBUTED PAPERS: MEASUREMENT ERROR AND ROBUST ESTIMATION

Chair: Jianxuan Liu, Syracuse University

8:30-8:45 AM

Integrating Segmentation Uncertainty into Statistical Analyses of Brain Volumes

Christina Chen, University of Pennsylvania

Multi-atlas image segmentation is a widely used approach in imaging studies that involve, for example, estimating the volume of a region of interest (ROI). However, current practices typically treat these images equally without incorporating the fact that the registration quality might vary among subjects. We propose a method that estimates the variance of the ROI volume estimate for each subject due to the multi-atlas segmentation procedure and thus provides a way of reweighting these estimates to increase efficiency in downstream estimation problems.

8:45-9:00 AM

Removal of Batch Effects in the Presence of Outliers via Robust ComBat

Andrew Chen, University of Pennsylvania

To collect larger samples in neuroimaging, many consortia have launched multi-center studies acquiring images from multiple scanners. These studies are known to suffer from scanner-related biases called scanner effects, which can obscure important biological associations and potentially drive spurious findings. Particularly in these large studies, outliers are often present in the data due to variations introduced in image acquisition and preprocessing. The state-of-the-art harmonization method called ComBat has been applied to deal with scanner effects in the mean and variance of features. However, the estimators of mean and variance used in ComBat are known to be sensitive to outliers and ComBat has yet to be evaluated for its robustness. We show in a multi-center study that outliers severely bias ComBat estimation, leading to major issues in the location and scale estimates of harmonized data. We propose a novel method using robust estimators in the ComBat framework. We demonstrate that our method yields harmonized multi-center data without biases driven by outliers. We further evaluate the effectiveness of our method in data with synthetic outliers and simulations.

9:00-9:15 AM

Application of the MC-SIMEX Method to a Weibull Accelerated Failure Time Model

Varadan Sevilimedu, Memorial Sloan Kettering Cancer Center

Misclassification of binary covariates is pervasive in clinical data and can lead to biased parameter estimates. Even though the effect of misclassification has been extensively studied in Cox proportional hazards models, its impact on Weibull accelerated failure time (AFT) models has not been studied. In this analysis, we study the bias caused by misclassification in binary covariates in a Weibull AFT model and explore the use of the misclassification simulation extrapolation method (MC-SIMEX) in correcting for this bias, along with its asymptotic properties. Simulation studies are carried out to investigate the numerical properties of the resulting estimator. The proposed method is then applied to colon cancer data obtained from the cancer registry at Memorial Sloan Kettering Cancer Center (MSKCC).

9:15-9:30 AM

High-Dimensional Measurement Error Models for Lipschitz Loss Functions

Xin Ma, Florida State University

Recently emerged biomedical data pose exciting opportunities for scientific discoveries. However, the ultrahigh dimensionality and non-negligible measurement errors of the data features create potential difficulties for statistical estimation and feature selection. There are limited existing measurement error models involving high-dimensional covariates, which usually require knowledge of the noise distribution and typically focus on linear or generalized linear models. In this work, we extend the high-dimensional measurement error models to a broader class of loss functions with Lipschitz continuity without requirement of the noise distribution. We subsequently propose a Lasso analog version of the method that is computationally scalable to much higher dimensions. We derive theoretical guarantees even when the number of covariates increases much faster than the sample size. Extensive simulation studies demonstrate superior performance compared to existing methods in classification and quantile regression problems. We apply the approach to a gender classification task based on functional connectivity and identify significant network edges that reveal gender differences.

9:30-9:45 AM

Misclassification Adjusting Supervised Machine Learning Algorithm

Eunchan Bae, University of Pennsylvania

Recent advancement in machine learning facilitates a deeper understanding of biomedical research. Automatic segmentation in biomedical imaging is one of the areas that has flourished with machine learning. Most supervised machine learning algorithms rely on assumptions that gold standard manual labels and measurements are accurate. However, if the labels or measurements are inaccurate, supervised algorithms become unreliable. In biomedical imaging, misclassification of labels is common due to inhomogeneous intensities in images, low-resolution images, and manual segmentation variability. Therefore, there is a need to relax the assumptions of no misclassification when building supervised machine learning algorithms. While several models exist to adjust the measurement errors, few models exist to adjust the misclassification. Here, we propose a novel iterative misclassification-adjusting supervised machine learning algorithm (ITEMS) that estimates the false-positives rates and false-negatives rates of the error-prone labels, self-corrects the labels, and conducts less biased estimation.

9:45-10:00 AM

Unified Robust Estimation

Zhu Wang, The University of Tennessee Health Science Center

Robust estimation is primarily concerned with how to provide reliable parameter estimates in the presence of outliers. Numerous robust loss functions have been proposed in regression and classification, along with various computing algorithms. In modern penalized generalized linear model (GLM), however, there is limited research on robust estimation that can provide weights to determine outlier status of the observations. This article proposes a unified framework based on a large family of loss functions, a composite of concave and convex functions (CC-family). Properties of the CC-family are investigated, and CC-estimation is innovatively conducted via the iteratively reweighted convex optimization (IRCO), a generalization of the iteratively reweighted least squares in robust linear regression. For robust GLM, the IRCO becomes iteratively reweighted GLM. The unified framework contains penalized estimation and robust support vector machine and is demonstrated with a variety of data applications.

10:00-10:15 AM

Accurate Confidence and Bayesian Interval Estimation for Non-Centrality Parameters and Effect Size Indices

Kaidi Kang, Vanderbilt University

Reporting effect size index estimates with their confidence intervals (CIs) can be an excellent way to simultaneously communicate the strength and precision of the observed evidence. We recently proposed a robust effect size index (RESI) that is advantageous over common indices because it is widely applicable to different types of data. Here, we use statistical theory and simulations to develop and evaluate RESI estimators and confidence/credible intervals that rely on different covariance estimators. Our results show (1) counter to intuition, the randomness of covariates reduces coverage for Chi-squared and F CIs; (2) when the variance of the estimators is estimated, the non-central Chi-squared and F CIs using the parametric and robust RESI estimators fail to cover the true effect size at nominal level. Using the robust estimator along with the proposed non-parametric bootstrap or Bayesian (credible) intervals provides valid inference for RESI, especially when model assumptions may be violated. We propose a framework for the analysis of effect size, such that effect sizes with confidence/credible intervals can be easily reported in an analysis of variance (ANOVA) table format.

56. CONTRIBUTED PAPERS: LONGITUDINAL DATA AND METHODS FOR BIOMEDICAL RESEARCH

Chair: Mark Meyer, Georgetown University

8:30-8:45 AM

Data Collected from Digital Health Tools: Statistical Challenges and Opportunities

Paramita Saha Chaudhuri, Biogen

Data collected from digital health tools such as smartphones, tablets, or wearable sensors are increasingly common in clinical trials or real-world studies. Sensor-derived digital outcomes offer opportunities over traditional clinical outcomes, allowing measurements to capture disease states in the free-living environment at more frequent intervals than in clinical settings and opening the possibility to learn about new disease prognostic, diagnostic or treatment efficacy markers. However, designing studies with digital tools and analyzing data collected from such tools pose statistical challenges. We give an overview of five challenges. First, suboptimal adherence to digital tools and selective missing data in the free-living environment must be handled at both the study design and analysis phases. Second, digital data are high-dimensional because multiple sensors each generate several features sampled longitudinally at a high frequency. The last three challenges are around choosing how frequently to collect digital data, accounting for practice effects, and combining features into composite outcomes. We contextualize the challenges using examples in neurosciences.

8:45-9:00 AM

A Joint Normal-Binary (Probit) Model for High-Dimensional Data

Margaux Delporte, KU Leuven

In many biomedical studies multiple responses are collected over time. This results in high-dimensional longitudinal data. The continuous and binary responses can be modelled jointly with joint generalized mixed models, where the random effects are allowed to correlate. This allows examining the association between the responses and how this association evolves over time. Investigating the association between the responses is often limited to scrutinizing these latent correlations between the random effects. This approach is extended by deriving closed-form formulas for the manifest correlations, which are the correlations between the observed responses. Next, the marginal joint model is constructed where the interpretation is no longer conditional on the random effects. From the marginal model, conditional models are derived. These lend themselves to predictions of a subvector of one response conditional on subvectors of the other responses and a subvector of the predicted response. Corresponding prediction and confidence intervals are constructed. Two case studies are discussed, in which pseudo-likelihood methodology is applied to reduce computational complexity.

9:00-9:15 AM

Bivariate Mixed Effects Model with Non-Stationary Stochastic Processes for Prediction of Rapid Disease Progression

Ziyun Wang, Cincinnati Children's Hospital Medical Center and University of Cincinnati

There are often multiple, related, noisily-measured outcomes that are critical to monitoring and predicting disease progression of individuals over time. Recent breakthroughs in real-time prediction have been achieved by replacing the classic random slope in the linear mixed effects model with a more flexible term representing a non-stationary stochastic process. The resulting model has been used to form predictive probabilities for clinically relevant target functions involving rates of change in the mean response function. However, this approach has been limited to a single outcome. Considering the case of two outcomes, we propose a bivariate mixed effects model utilizing integrated Brownian motion for each mean response function. The proposed bivariate target function simultaneously predicts under the two-outcome scenario based on clinically meaningful thresholds of rates of change. This novel approach is applied to achieve real-time prediction of key changes in nutrition and lung function for children with cystic fibrosis who are followed in a national patient registry.

9:15-9:30 AM

Likelihood-Based Inference for Skewed Responses in a Crossover Trial Setup

Savita Pareek, IIT Bombay, India

This work proposes a statistical model for crossover trials with multiple skewed responses measured in each period. A 3*3 crossover trial data where different doses of a drug were administered to subjects with a history of seasonal asthma rhinitis to grass pollen is used for motivation. In each period, gene expression values for ten genes were measured from each subject. It considers a linear mixed effect model with skew normally distributed random effect or random error term to model the asymmetric responses in the crossover trials. The paper examines cases (i) when a random effect follows a skew-normal distribution, as well as (ii) when a random error follows a skew-normal distribution. The EM algorithm is used in both cases to calculate maximum likelihood estimates of parameters. Simulations and crossover data from the gene expression study illustrate the approach.

9:30-9:45 AM

Flexible Model to Estimate Patient-Specific Timing and Degree of Rapid Decline in Alzheimer's Disease (AD)

Mohammad Bhuiyan, LSU Health Shreveport

Alzheimer's disease (AD) is a progressive, genetic disease characterized by frequent, prolonged drops in cognitive function. Accurately predicting rapid cognitive-function decline is essential for clinical decision support and timely intervention. Determining whether an individual is experiencing a period of rapid decline is complicated due to its heterogeneous timing and extent, and the error component of the measured brain function. The most common approach is conventional linear mixed modeling-estimating a population-level slope of brain function decline and using random effects to address serial correlation-but this ignores nonlinear features of disease progression and distinct sources of variability. We propose a flexible model to estimate patient-specific timing and degree of rapid decline while appropriately characterizing natural progression and variation in AD.

9:45-10:00 AM

Statistical Methods for Identifying Time-Varying Genetic Effects in Longitudinal Genetic Studies

Amei Amei, University of Nevada, Las Vegas

Many genetic studies contain rich information on longitudinal phenotypes that require powerful analytical tools for optimal analysis. Genetic analysis of longitudinal data that incorporates temporal variation is important for understanding the genetic architecture and biological variation of complex diseases. Most of the existing methods assume that the contribution of genetic variants is constant over time and fails to capture the dynamic pattern of disease progression. Here, we propose a retrospective varying coefficient mixed model association test, RVMMAT, to detect time-varying genetic effect on longitudinal binary traits. We model dynamic genetic effect using smoothing splines, estimate model parameters by maximizing a double penalized quasi-likelihood function, and evaluate statistical significance via a retrospective approach to achieve robustness to model misspecification. We applied RVMMAT to a genome-wide association analysis of longitudinal measure of hypertension in the Multi-Ethnic Study of Atherosclerosis.

10:00-10:15 AM

Regression Approaches to Assess Effect of Treatments that Arrest Progression of Symptoms

Ana Maria Ortega-Villa, NIAID/NIH

Randomized controlled trials are known as the gold-standard to measure the effects of an intervention or treatment. However, there are cases in which randomizing a participant to placebo is unethical, e.g. cases where the progression of disease is irreversible, like in neonatal-onset multisystem inflammatory disease (NOMID). Withholding treatment for NOMID patients is detrimental to patient health. In this work, we propose to estimate the effect of treatment using a bent line model. This model estimates the longitudinal trajectory of the outcome variable and adds an interaction term between a treatment indicator and the time since treatment initiation. The interaction parameter allows us to determine whether there is a change in the slope associated with treatment. This method is appropriate for cases in which treatment slows or arrests the effect of the disease on the outcome, but does not cure it, as is the case in NOMID. We evaluate the performance of the bent line and other alternative approaches via simulation and illustrate the methodology through a prospective cohort of NOMID patients enrolled at the NIH clinical center.

57. CONTRIBUTED PAPERS: PERSONALIZED MEDICINE

Chair: Gege Gui, Johns Hopkins University

8:30-8:45 AM

Optimal When-to-Treat Policies Under Dynamic Resource Constraints

Tarek Zikry, University of North Carolina at Chapel Hill

Precision medicine uses data to leverage patient heterogeneity to learn optimal tailored treatments. While many strategies have been proposed for learning optimal tailored treatment policies, only recently have methods been developed for dealing with the question of treatment timing and separately, attempts have been made to incorporate resource constraints. Using the when-to-treat (WTT) policy class introduced by Nie, Brunskill, and Wager (2021), we propose a policy learning algorithm to find an optimal WTT policy under a dynamic resource constraint in an indefinite time horizon setting. Crucially, our work accounts for the impact of allocating resources to patients today on the health and disease progression for patients treated tomorrow. We compare our algorithm to na?ve resource constraint-free WTT policies in addition to other standard policy learning algorithms for dynamic treatment regimes, and assess performance in both simulated and real-world resource-constrained data.

8:45-9:00 AM

BaySyn: Bayesian Evidence Synthesis for Multi-system Multiomic Integration

Rupam Bhattacharyya, University of Michigan

The discovery of cancer drivers/drug targets are often limited to biological systems, e.g., cancer models/patients. While multiomic patient databases have sparse drug response, model systems databases provide lower lineage-specific sample sizes, resulting in reduced power to detect functional drivers and their associations with drug sensitivity. Hence, integrating evidence across model systems can more efficiently deconvolve cancer cellular mechanisms and learn therapeutic associations. To this end, we propose BaySyn - a hierarchical Bayesian evidence synthesis framework for multi-system multiomic integration. BaySyn detects functional driver genes based on their associations with upstream regulators and uses this evidence to calibrate Bayesian variable selection models in the outcome layer. We apply BaySyn to multiomic datasets from CCLE and TCGA, across pan-gynecological cancers. Our mechanistic models implicate several functional genes such as PTPN6 and ERBB2 in the KEGG adherens junction gene set. Further, our outcome model makes more discoveries in drug response models than uncalibrated models at similar Type I error control - such as BCL11A (breast) and FGFRL1 (ovary).

9:00-9:15 AM

Clustering Matrices: A Metric Learning Approach to Disease Subtyping in Mental Health

Hanchao Zhang, Grossman School of Medicine, New York University

Multiple clustering algorithms have been developed for subtyping diseases, especially for mental disorders that are not very clearly defined. Most of the developed methods focus on clustering scalar outcomes or functional outcomes. This talk focuses on the development of clustering algorithms for matrix-valued data that are encountered more and more frequently in practice (e.g., functional connectivity matrices). In order to cluster matrices, a distance metric is needed, and in this talk, a metric learning approach for clustering is investigated in order to optimize clustering procedures. Specifically, we will present results comparing matrix-clustering based on metric learning comparing distance metrics for Stiefel manifolds and ellipsoids as well as common principal components (Flury, 1987). This clustering aims to identify possible subtypes of diseases using matrix-valued data guided by diagnostic labels and will be particularly useful for data arising from psychiatric illnesses where disease severity can vary over a continuous spectrum.

9:15-9:30 AM

A First Step Toward Precision Medicine: Prior Knowledge-assisted Integrative Convex Clustering for Disease Subtyping

Xiaoyu Zhang, Department of Biostatistics, Boston University School of Public Health

Accurate disease subtyping could be the first step toward precision medicine which aims to provide the right treatment for the right group of patients. Multi-omics data and accumulated biological knowledge may offer a great opportunity to explore diseases mechanisms and subtypes. However, the existing clustering methods, such as Sparse Convex Clustering (SCC), cannot directly utilize the prior knowledge even though SCC produces stable clusters. We develop a novel clustering method, Prior Knowledge-assisted Integrative Convex Clustering (PK-ICC), to respond to the need for disease subtyping in precision medicine. We incorporate prior biological knowledge such as pathways of genes or metabolites, cis-regulatory mechanism between gene expression, methylation, and copy number variation through a group lasso penalty to improve disease subtyping and select biological meaningful groups of features simultaneously. We conduct simulation studies under scenarios with various prior knowledge accuracy to evaluate the performance of our method. We also illustrate our method with mRNA expression data of lung cancer for disease subtyping and important biomarkers identification.

9:30-9:45 AM

A New Concept of Individual Reference Intervals and Individual Reference Change Values through Joint Quantile Models

Murih Pusparum, Data Science Institute, I-Biostat, Hasselt University, Belgium; Flemish Institute for Technological Research (VITO)

A population reference interval (RI) for clinical tests (e.g. HDL cholesterol) describes lower and upper bounds, so it contains 95% of the outcomes in a healthy population. They are used to compare a new test result: when it falls within the RI, the medical practitioner would indicate a healthy-normal reading. However, an RI only gives a single interval to be used for all subjects, lacking precise interpretation in an individual context. A new concept of Individual Reference Interval (IRI) has been proposed as a valid alternative, relevant to precision and personalized health landscapes. To obtain the correct estimates of IRIs, we have developed quantile models in combination with penalization methods that allow for information-sharing among subjects so that IRIs can be computed with only short time series for each subject. Moreover, we also introduce the concept of Individual Reference Change Values (I-RCV) for detecting a sudden change between two measurements. We present simulation studies for evaluating the new methods, as well as an illustration with real-life longitudinal data collected by accredited laboratories.

9:45-10:00 AM

Estimating Heterogeneous Survival Treatment Effect under Counterfactual Framework

Na Bo, University of Pittsburgh

Estimating heterogeneous treatment effect plays a central role in personalized medicine as it provides critical information for tailoring existing therapies for each patient to get the optimal treatment. Recently, meta-learning approaches have received a lot of attention in estimating conditional average treatment effect (CATE) by using multi-step algorithms coupled with flexible machine learning methods. In this project, we provide a meta-learning framework to estimate CATE on survival outcomes. We consider several pseudo-CATE regression approaches along with popular machine learning methods such as random survival forests, Cox-Lasso, and survival neural networks. We address advantages and challenges in implementing these methods to survival outcomes through comprehensive simulations and provide guidelines for applying these methods to survival outcomes in different situations. Finally, we demonstrate the methods by analyzing a large randomized clinical trial, the AREDS study for an eye disease, age-related macular degeneration, to estimate CATE and make individualized treatment recommendations.

10:00-10:15 AM

A Joint Model for Time-to-Event Data with Heterogeneity in Response to Treatments

Qiao Zhang, NYU Grossman School of Medicine

Heterogeneity in treatment outcomes is common for complex human diseases due to drug resistance and other differential responses to treatments. The paradigm of modern medicine focuses more on personalized therapies. It is critical to identify homogeneous subgroups of patients to develop more efficacious targeted therapies or optimize individualized treatment decisions. A group-specific model that accounts for heterogeneity often has high predictive accuracy in precision medicine, even when the group indicator is partially missing. We develop a joint model for two time-to-event outcomes with random censoring and propose an EM algorithm to maximize the joint model likelihood. A penalized likelihood method with adaptive LASSO penalty is used in the algorithm for variable selection. Logistic regression is used to link the two time-to-event models and facilitates identification of the latent classes. The resulting model can provide a better estimation of personalized patient survival than the conventional Cox proportional hazard models. The effectivity of the model is evaluated by numerical studies and applied to the National Alzheimer’s Coordinating Center (NACC) dataset.

58. CONTRIBUTED PAPERS: MISSING DATA METHODS AND APPLICATIONS

Chair: Sedigheh Mirzaei, St. Jude Children’s Research Hospital

8:30-8:45 AM

Statistical Methods for Modeling Exposure Variables Subject to Limit of Detection

Eunsil Seok, Department of Population Health, New York University Grossman School of Medicine

In environmental health research, it is of great interest and critical importance to evaluate the effect of environmental exposures. One of the common issues with measuring exposures is that values below the laboratory limit of detection (LOD) are often not detected. Inappropriate handling of missing due to LOD may lead to erroneous scientific results directly related to individual and population health. Various statistical models and approaches have been proposed to address this problem. In this work, we examine and compare various methods of handling exposure variables subject to LOD as covariates or outcome in a model and evaluate the performance of methods under different scenarios. Different methods including complete case analysis, fill-in method, multiple imputation, missing indicator model, two-part model, Tobit model, and many others, are illustrated through a dataset from NHANES 2013-2014 survey including 3 chemical exposures. Our results and recommendations can serve as a guide for modeling exposure variables in future epidemiology research.

8:45-9:00 AM

Data Fusion for Time to Event Outcomes

Fatema Shafie Khorassani, University of Michigan, School of Public Health

Data fusion is a particularly challenging scenario in data integration in which the probability of observing complete data is zero for every subject. The goal is to make inference about a model regressing an outcome on covariates coming from two separate sources. The outcome of interest is collected in one dataset, and a set of variables is collected in another. Both datasets collect a common subset of variables. We propose a method for data fusion with a time-to-event outcome by applying a proportional hazards model and transforming the observed datasets to apply an equivalent Poisson model in order to derive the appropriate semiparametric estimating equations for data fusion. The class of semiparametric estimating equations includes a doubly robust (DR) equation which provides consistent parameter estimates if either the data source process or the distribution of unobserved covariates is correctly specified. We apply the estimating equations to studying racial disparities in cancer specific mortality data from the National Cancer Institute?s Surveillance, Epidemiology, and End Results registry adjusted for confounders collected in the National Cancer Database.

9:00-9:15 AM

An Approximated Expectation-Maximization (EM) Algorithm for Integrative Analysis of Datasets with Nonresponse

Jiahe Li, University of Pittsburgh

Missing data are pervasive in public health studies. Standard statistical methods often require unverifiable assumptions and modelling of the missing-data mechanism. Misspecification of missingness models often leads to biased estimates and wrong conclusions. For integrative analysis of data from multiple studies, the issue with missingness is exacerbated that the missing process varies. Modelling study-specific missingness under a unified framework is prohibitive. Here we propose an approximated expectation-maximization (AEM) algorithm for the integrative regression analysis of data with nonresponse assuming the datasets follow the same regression model and are independent. Each dataset may suffer from an arbitrary missing mechanism. With a consistent initial estimator from a prior study or a complete dataset, the AEM algorithm avoids modelling missing mechanisms and yields a more efficient estimator. Simulation studies are used to illustrate the efficiency gain under various settings.

9:15-9:30 AM

Imputing and Summarizing Multiple Correlated High-Frequency Digital Biomarkers

Nicole Wakim, University of Michigan

High-frequency digital biomarkers can improve the detection and prediction of human disorders and diseases. However, it is challenging to leverage biomarkers in traditional modeling techniques as the outcome of interest is often assessed less frequently than the biomarkers. In addition, biomarkers rarely are completely observed. Thus, it is necessary to summarize biomarker data to match the outcome's frequency while simultaneously imputing missing biomarker values. We investigate the use of multiple biomarkers to impute missing values and their effect on downstream association analysis with a longitudinal outcome. We propose a method using linear mixed effects models (LMM) incorporating correlation between biomarkers to impute missing values. Via simulation, we assess our method across scenarios with increasing numbers of biomarkers and different levels of correlation in the context of longitudinal analysis of a binary outcome. Our results show that the LMM imputation method is more efficient than available data and less biased than other imputation methods. A real data analysis involving mild cognitive impairment is used to demonstrate the empirical differences.

9:30-9:45 AM

Comparison of Reference-Based Multiple Imputation Methods with Repeated Sampling Variance Estimator in Longitudinal Clinical Trials

Yusuke Yamaguchi, Astellas Pharma Global Development Inc.

Reference-based multiple imputation (RBMI) has gained popularity to be used as primary analysis in longitudinal clinical trials when the estimand follows a treatment policy strategy for handling treatment discontinuation. Rubin's variance estimator is well known to be biased when it is used along with the RBMI due to uncongeniality, which can yield an overly conservative treatment effect estimate. Alternatively, several approaches to estimate a repeated sampling variance have been proposed in the context of RBMI, such as congenial Bayesian estimators and bootstrap estimators. The repeated sampling variance estimator has attractive features especially in the case that the RBMI is used as the primary analysis; however, there has been little investigation as to relative performance of the methods. We conducted an extensive simulation study to examine the comparative performance of several repeated sampling variance estimators under three variants of RBMI, including jump to reference, copy reference, and copy increments in reference. The simulation study revealed operating characteristics of each variance estimator, which would guide the appropriate implementation of the RBMI.

9:45-10:00 AM

Semiparametric Estimation of Misclassified Semi-Competing Risks Data Under Gamma-Frailty Conditional Markov Model

Ruiqian Wu, University of Nebraska Medical Center

Semi-competing risks data model has become increasingly popular in studying the association between times to disease progression and death, and allows us to better understand how an intermediate event impacts the terminal event. However, in many applications event ascertainment is incomplete resulting in event misclassification that complicates the statistical inference based on semi-competing risk data models. In this work, we consider a Gamma frailty conditional Markov model to study the misclassified semi-competing risk data, and propose a two-stage semiparametric maximum pseudo-likelihood estimation approach equipped with a pseudo-EM algorithm to make unbiased statistical inference, in which the probability of event misclassification is estimated via a nonparametric regression splines estimation procedure in the first stage. Extensive simulation studies show the proposed method is numerical stable and performs well even under a large amount of event misclassification. The method is applied to a multi-center HIV cohort study in East African to measure the impact of interruption of lifelong antiretroviral therapy (ART) on HIV mortality.

10:00-10:15 AM

Analysis of Methods for Handling Left-Censored Observations

Clifford Crafford, University of Memphis, St. Jude Children's Research Hospital

When analyzing biomarkers, the analytes of interest may be subject to the limitations of the tools of measurements. In such cases, some assays cannot be accurately reported as some of the quantity may fall below minimum observable levels. In these circumstances the assay is said to be Below the Level of Detection (BLoD) or Left-Censored. Many statistical methods exist for analyzing data subject to such limits. However, most of these methods have not been evaluated in circumstances involving a case-control experiments where the case and control groups possess different levels of censoring of the biomarker. The work present here seeks to evaluate six conventional methods on their performance for detecting the difference in two population means where both populations involve BLoD biomarker with different levels of censoring. Simulation study approaches are used to estimate the power in each method under different assumptions about the distributions of the biomarkers. These include the use of common methods, such as; the Simple Imputation, Beta Imputation, Multiple Imputation, Maximum Likelihood, Mixture Modeling, and the Cox Regression.

59. CONTRIBUTED PAPERS: VARIABLE SUBSET SELECTION/MODEL SELECTION

Chair: Elizabeth Chase, University of Michigan

8:30-8:45 AM

Combine Immune Responses to Study Heterogeneous Infectious Risk in the Immune Correlates Analysis of HIV Vaccine Studies

Chaeryon Kang, Dept. of Biostatistics, University of Pittsburgh

In HIV vaccine/prevention research, probing into the vaccine-induced immune responses that can help predict the risk of HIV infection provide valuable information for developing vaccine regimens. Previous correlate analysis of the Thai vaccine trial aided the discovery of interesting immune correlates related to the risk of developing an HIV infection. The present study aimed to identify the combinations of immune responses associated with the heterogeneous infection risk. We explored a “change-plane” via a combination of a subset of immune responses that could help separate vaccine recipients into two heterogeneous subgroups regarding the association between immune responses and the risk of developing an infection. Additionally, we developed a new variable selection algorithm through a penalized likelihood approach to investigate a parsimonious marker combination for the change-plane. The application of the proposed statistical approach to the Thai trial has been presented, wherein the marker combinations were explored among several immune responses and antigens.

8:45-9:00 AM

Penalized Bayesian Forward Continuation Ratio Model with Application to High-dimensional Data with a Discrete Survival Outcome

Anna Seffernick, St. Jude Children's Research Hospital

While time-to-event data are often continuous, there are several instances where discrete survival data, which is inherently ordinal, might be more appropriate or useful. Several discrete survival models exist, but the forward continuation ratio (FCR) model with complementary log-log (clog-log) link has a survival interpretation and is closely related to the Cox proportional hazards model, despite being an ordinal model. This FCR model has previously been implemented in the high-dimensional setting using the ordinal generalized monotone incremental forward stagewise (OGMIFS) algorithm. Here, we propose a Bayesian penalized FCR model with clog-log link and explore different priors to perform variable selection and regularization. Simulation studies are used to compare the variable selection performance of these priors, as well as to compare our method with existing methods. We also illustrate our model on acute myeloid leukemia omics datasets to identify proteomic and genomic features associated with discrete survival.

9:00-9:15 AM

Tree-Guided Rare Feature Selection and Logic Aggregation with Electronic Health Records Data

Jianmin Chen, Department of Statistics, University of Connecticut

Statistical learning with a large number of rare binary features is commonly encountered in analyzing electronic health records (EHR) data, especially in the modeling of disease onset with medical diagnoses and procedures. Dealing with the sparse and binary feature matrix is challenging as conventional methods may suffer from a lack of power in testing and inconsistency in modeling while machine learning methods may not produce interpretable results. To improve EHR-based modeling and use the hierarchical structure of disease classification, we propose a tree-guided feature selection and logic aggregation approach for regression with rare binary features, in which dimension reduction is achieved through not only a sparsity pursuit but also an aggregation promoter with the logic operator of "or". We convert the combinatorial problem into a convex linearly-constrained regularized estimation with theoretical guarantees. In a suicide risk study, our approach is able to select and aggregate mental health diagnoses as guided by the hierarchy of the International Classification of Diseases, and improves both prediction and interpretation.

9:15-9:30 AM

A Fast Solution to the Lasso Problem with Equality Constraints

Lam Tran, University of Michigan Department of Biostatistics

The equality-constrained lasso problem augments the standard lasso by imposing equality constraints on regression coefficients. Due to the inseparability of predictors in the constrained lasso, highly optimized approaches to solve the standard lasso are unable to be used off-the-shelf. Existing constrained lasso algorithms are computationally inefficient and have only been applied to linear and logistic models. To address these limitations, we propose an efficient algorithm by combining two acceleration schemes: a candidate subset approach and a two-stage optimization approach. The proposed method is several magnitudes faster than existing constrained lasso fitters, while resulting in identical solution paths. It can also be easily adapted to time-to-event outcomes. We demonstrate the efficacy of our method on simulations and real data examples, including a log-contrast model for the oral microbiome and a gene-pair selection model on myeloma RNA-seq data. In these real data examples, our method accounts for additional unpenalized and unconstrained covariates as well as candidate predictor dimensions in the thousands, which existing methods are unable to accommodate.

9:30-9:45 AM

A Convex-Nonconvex Strategy for Grouped Variable Selection

Xiaoqian Liu, The University of Texas MD Anderson Cancer Center

This paper deals with the grouped variable selection problem. A widely used strategy is to augment the negative log-likelihood with a sparsity-promoting penalty. Existing methods include the group Lasso, group SCAD, and group MCP. The group Lasso solves a convex optimization problem but is plagued by underestimation bias. The group SCAD and group MCP avoid this estimation bias but require solving a nonconvex optimization problem that may suffer from suboptimal local optima. In this work, we propose an alternative method based on the generalized minimax concave (GMC) penalty, which is a folded concave penalty that maintains the convexity of the objective function. We develop a new method for grouped variable selection, the group GMC, that generalizes the original GMC method. We present an efficient algorithm for computing the group GMC estimator and prove properties of the solution path to guide its numerical computation and tuning parameter selection. We establish error bounds for both the group GMC and original GMC estimators. A rich set of simulation studies and a real data application indicate that group GMC outperforms existing methods under a wide array of scenarios.

9:45-10:00 AM

A Clustering Approach with Variable Selection for Longitudinal Data

Marie Denis, Cirad/Georgetown University

In medicine or agronomy, longitudinal studies are conducted to understand dynamic processes, such as disease progress or growth. The identification of groups of individuals with similar profiles over time along with their associated genetic markers can help in the development of more effective therapeutic strategies in human disease and in gaining insights into the adaptation of plants to climate change. Most existing statistical methods do not allow the simultaneous analysis of longitudinal outcomes and the selection of relevant markers in high-dimensional data. In this talk, I will present a Bayesian approach that combines mixture of mixed effects models and variable selection to identify groups of individuals with similar longitudinal response profiles and their associated subsets of covariates with time varying effects. I will illustrate the performance of the approach with simulated data and yeast cell-cycle gene expression data.

10:00-10:15 AM

A Sparse Multivariate Regression Approach for Estimating Covariance Matrices with Covariates

Rakheon Kim, Baylor University

In multivariate analysis, detecting the linear relationships among variables is important and such linear relationships are encoded in the covariance matrix. However, in some applications, the covariance structure may depend on each individual's characteristic. In this paper, we consider estimation of subject-specific covariance matrices by considering the effects of covariates to the covariance matrices. We cast this problem as the multivariate linear regression for the covariance elements in the matrix and propose regularization approaches to address the complexity of the model. Also, for individual modeling of the variance elements, we propose to minimize the entropy loss of our estimator given that the off-diagonal elements have already been fixed by the multivariate regression. Our algorithm for this diagonal estimation ensures the estimator for each individual to be positive definite. We study the benefits of estimating the covariance matrix with covariates in simulated settings and apply the method to the real data to detect subject-specific covariance network for a Parkinson's disease dataset.

60. CONTRIBUTED PAPERS: NEW METHODS IN STATISTICAL GENETICS AND GENOMICS

Chair: Wenan Chen, St. Jude Children’s Research Hospital

8:30-8:45 AM

Identification and Inference for High-Dimensional Pleiotropic Variants in GWAS

Lap Sum Chan, University of Michigan

Current statistical methods for identifying pleiotropic variants in genome-wide association studies (GWAS) are based on two-step approaches, are susceptible to spurious discoveries. We propose a new statistical approach, termed Debiased-regularized Factor Analysis Regression Model (DrFARM), through a joint regression model for a simultaneous analysis of high-dimensional genetic variants and multilevel dependencies, to identify pleiotropic variants in multi-trait GWAS. This joint modeling strategy fosters the capacity of controlling an overall error so to permit a universal false discovery rate (FDR) control. This methodology utilizes the strengths of the debiasing technique and the Cauchy combination test, both being theoretically justified, to establish a valid post-variable selection inference on pleiotropic variants. Through extensive simulations, we show that DrFARM can appropriately control the overall FDR. Applying DrFARM on 1,031 metabolites measured on 6,135 men from the Metabolic Syndrome in Men (METSIM) study, we identify new pleiotropic loci for 16 metabolite pairs at p < 7.2 x 10-11.

8:45-9:00 AM

A Flexible Zero-Inflated Poisson-Gamma Model with Application to Microbiome Sequence Count Data

Roulan Jiang, Tsinghua University

In microbiome studies, it is of interest to use a sample of microbes to estimate the population proportion of these taxa. However, due to biases introduced in sampling and preprocessing steps, these observed taxa abundances may not reflect true patterns. Repeated measures including longitudinal study are potential solutions to mitigate the discrepancy between observations and true underlying abundances. Yet, widely observed zero-inflation and over-dispersion issues can distort downstream statistical analyses aiming to associate taxa abundances with covariates of interest. To address the aforementioned challenges, we propose a Zero-Inflated Poisson Gamma (ZIPG) framework. From a perspective of measurement errors, we accommodate the discrepancy between observations and truths by decomposing the mean parameter into a true abundance and a multiplicative measurement of sampling variability. We connect mean abundance and variability to different covariates, and build valid statistical inference procedures. Through comprehensive simulation and real data analysis, the proposed ZIPG method provides significant insights into distinguished differential variability and abundance.

9:00-9:15 AM

Heritability Estimation with High Efficiency using Linkage Disequilibrium and GWAS Summary Statistics (HEELS)

Hui Li, Harvard T.H. Chan School of Public Health

Heritability, defined as the proportion of variability in the phenotype that is attributable to genetic factors, is a fundamental parameter in statistical genetics. Many existing heritability estimation methods have been developed in the framework of restricted maximum likelihood (REML) method under the linear mixed models using data from genome-wide association studies (GWAS), but the requirement to access individual-level data severely limits their applicability. Alternatively, methods that rely on GWAS summary-statistics can be applied more broadly, but yield less statistically efficient estimators. We introduce a new estimator that has comparable variance as REML but only requires summary-level data. The relative efficiency (RE) of our estimator compared to REML is as high as 92% whereas other state-of-the-art summary-statistics-based methods, such as LDSC and GRE, have a RE below 25%. We demonstrate the statistical efficiency of our estimator and the advantages of the proposed sparse representation of the LD matrix through both simulations and empirical analyses in the UK Biobank.

9:15-9:30 AM

A Scalable Statistical Framework for Genome-Wide Interaction Testing Harnessing Cross-Trait Correlations with an Application to Alzheimer's Disease

Shijia Bian, Department of Biostatistics and Bioinformatics, Emory University

We propose a scalable method for interaction testing using variance-based approaches but instead leveraging valuable information contained within multiple correlated phenotypes. We can formally show that SNPs with interactive effects yield differential correlation patterns among phenotypes per genotype category. Our proposed test first applies linear regression to assess the relationship between SNP genotype and pairwise cross-products among phenotypes. We then combine the resulting pairwise cross-product regression p-values together using an aggregated Cauchy statistic to form an optimal test. Our method, which we call SCAMPI, is computationally scalable to genome-wide analyses (with similar run times to variance-based interaction methods), can handle many phenotypes and can adjust for confounders. Type I error, and power simulation verify that SCAMPI is well calibrated and can detect even sparse interaction effects observed among a large group of modeled phenotypes. We further applied SCAMPI to - ROS/MAP study, an Alzheimer?s Disease study. SCAMPI is also scalable to large biobank data.

9:30-9:45 AM

Gene-Level Association Analysis of Bi-Variate Ordinal Traits with Functional Regressions

Shuqi Wang, Department of Biostatistics, Bioinformatics, and Biomathematics, Georgetown University Medical Center

In genetic studies, many phenotypes have multiple naturally ordered discrete values. Such phenotypes can be correlated to each other. If multiple correlated ordinal traits are analyzed simultaneously, the power of analysis may increase significantly. So, we propose bi-variate functional ordinal linear regression (BFOLR) models using latent regressions with cumulative logit link and probit link to perform gene-based analysis for bi-variate ordinal traits and sequencing data. In BFOLR models, genetic variant data are viewed as stochastic functions of physical positions, and genetic effects are treated as a function of physical positions. The correlation of two ordinal traits is considered via latent variables. BFOLR models are built on functional data analysis and can be revised to analyze bi-variate ordinal traits and high-dimension genetic data. Both rare and common variants can be analyzed. Simulation studies show the likelihood ratio test statistics of BFOLR models control type I errors well and have good power performance. BFOLR models are applied to Age-Related Eye Disease Study data. Two genes, CFH and ARMS2, are found to associate with macular degeneration strongly.

9:45-10:00 AM

A Pseudo-Value Regression Approach for Differential Network Analysis of Co-Expression Data

Seungjun Ahn, Department of Biostatistics, University of Florida

The differential network (DN) analysis detects changes in measures of association among genes under two or more conditions. In this study, we introduce a Pseudo-value Regression Approach for Network Analysis (PRANA). This is a novel method of DN analysis that also adjusts for additional clinical covariates. We start from mutual information (MI) criteria, followed by pseudo-value calculations, which are then entered into a robust regression model. Performance in terms of precision, recall, and F1 score of differentially connected (DC) genes is assessed in both univariable and multivariable settings through variety of simulations. By and large, PRANA outperformed dnapath and DINGO, neither of which is equipped to adjust for available covariates such as patient-age. Lastly, we employ PRANA in a real data application from the Gene Expression Omnibus (GEO) database to identify DC genes that are associated with chronic obstructive pulmonary disease (COPD) to demonstrate its utility. This is the first attempt of utilizing a regression modeling for DN analysis by collective gene expression levels between the two or more comparing groups with the inclusion of additional covariates.

10:00-10:15 AM

Two-stage Hypothesis Tests for Variable Interactions with FDR Control

Jingyi Duan*, Cornell University

In genome-wide association studies where dependences between variables commonly exist, it is often of interest to infer the interaction effects. However, testing pairwise interactions among millions of variables in high-dimensional data suffers from low statistical power and huge computational cost. Therefore, we propose a two-stage testing procedure with false discovery rate (FDR) control, which is known as a less conservative multiple-testing correction. The difficulty in the FDR control dues to the data dependence among test statistics and the number of hypothesis tests conducted in the second stage depends on the result in the first stage. By using the Cram?r type moderate deviation technique, we show that our procedure controls FDR at the desired level asymptotically in the GLM, where the model is allowed to be misspecified. In addition, the asymptotic power of the procedure is rigorously established. We show via comprehensive simulation studies that our procedure is computationally more efficient than the classical BH procedure with a comparable or improved statistical power. Finally, we apply the proposed method to a bladder cancer data from dbGaP.

Tuesday, March 21, 2023 | 1:45-3:30 PM

61. NOT ALL CENSORING IS A SURVIVAL PROBLEM: STATISTICAL METHODS FOR CENSORED COVARIATES

Organizer: Sarah Lotspeich, Wake Forest University
Chair: Kyle Grosser, University of North Carolina at Chapel Hill

1:45-2:10 PM

Use of Pseudo-Observations to Handle Censored Covariates

Jing Qian, University of Massachusetts, Amherst

We consider the problem of censored covariates, such as mother's age at onset of dementia, in regression models for Alzheimer's disease outcomes of offspring. In our prior work, we proposed threshold regression methods for testing of the significance of the effect of a censored covariate and for unbiased estimation of the regression coefficient of the censored covariate. Here we improve upon our thresholding approach by using all observations via insertion of pseud-observations for the survivor function, emulating the thresholding approach, and the mean survival time, emulating the original model. We correct bias through two-stage modeling. We conduct simulations to evaluate the finite-sample performance of proposed methods, and compare them with existing methods. We apply the methods to an Alzheimer's disease study.

2:10-2:35 PM

Regression with Interval-Censored Covariates: Application to Cross-Sectional Incidence Estimation

Douglas Morrison, University of California, Davis

A method for generalized linear regression with interval-censored covariates is described, extending previous approaches. A scenario is considered in which an interval-censored covariate of interest is defined as a function of other variables. Instead of directly modeling the distribution of the interval-censored covariate of interest, the distributions of the variables which determine that covariate are modeled, and the distribution of the covariate of interest is inferred indirectly. This approach leads to an estimation procedure using the Expectation-Maximization (EM) algorithm. The performance of this approach is compared to two alternative approaches, one in which the censoring interval midpoints are used as estimates of the censored covariate values, and another in which the censored values are multiply imputed using uniform distributions over the censoring intervals. A simulation framework is constructed to assess these methods' accuracies across a range of scenarios. The proposed approach is found to have less bias than midpoint analysis and uniform imputation, at the cost of small increases in standard error.

2:35-3:00 PM

Weighting Methods for Handling Censored Covariates in Huntington Disease Studies

Marissa Ashner, University of North Carolina at Chapel Hill

Censored covariates occur in data analysis when the true value of a regression covariate is unknown but is known to be greater than some value. While there are vast bodies of literature covering how to analyze missing data or censored outcomes, the literature base on censored covariates is small but growing. We construct and analyze methods that account for censored covariates in order to produce consistent and efficient estimators while making minimal assumptions about the underlying data structures. First, we clarify when estimators are guaranteed to be consistent for a complete case analysis and draw connections between various censoring mechanisms that lead to consistency. Secondly, we propose an augmented inverse probability weighting (AIPW) estimator, which weights each complete case to account for any selection bias and adds an augmentation term to improve upon efficiency. Finally, we will propose an augmented complete case estimator that combines the strengths of a complete case analysis and AIPW. All methods will be applied to the PREDICT-HD study to analyze psychiatric symptoms in Huntington disease patients is as a function of time to clinical diagnosis.

3:00-3:25 PM

It's Integral: Replacing the Trapezoidal Rule to Remove Bias and Correctly Impute Censored Covariates with Their Conditional Means

Sarah Lotspeich, Wake Forest University, Department of Statistical Sciences

Modeling symptom progression to prioritize subjects for a new Huntington's disease clinical trial is problematic since key covariate time to diagnosis can be censored. Imputation is an appealing strategy where censored covariates are replaced with their conditional means, but existing methods saw over 100% bias. Calculating these conditional means well requires estimating and integrating over the survival function of the censored covariate from the censored value to infinity. To flexibly estimate the survival function, existing methods use the Cox model with Breslow's estimator. Then, for integration, the trapezoidal rule, which is not designed for indefinite integrals, is used. This leads to bias. We propose a calculation that handles the indefinite integral with adaptive quadrature. Yet, even with adaptive quadrature, the integrand (the survival function) is undefined beyond the observed data. We identify the best method to extrapolate. In simulations, we show that replacing the trapezoidal rule with adaptive quadrature (plus extrapolation) corrects the bias. We further show how imputing with corrected conditional means helps prioritize patients for clinical trials.

62. ADVANCES IN STATISTICAL METHODS FOR NATIONAL DISEASE REGISTRIES

Organizer/Chair: Wenbo Wu, NYU Grossman School of Medicine

1:45-2:10 PM

Statistical Analysis of Clustered Semi-Continuous Data Truncated by Death with Application to Nursing Home Profiling

Sebastien Haneuse, Harvard T.H. Chan School of Public Health

Nursing homes provide residence and quality-of-life services for people who are aging and/or have physical and mental ailments. As the COVID-19 pandemic progressed, interest arose in studying nursing homes, in particular in relation to cost accrual and health care utilization. Investigating variation in these outcomes, however, requires consideration of a range of statistical intricacies including: the longitudinal nature of cost/utilization accrual over time; the clustering of patients within nursing homes; that the outcome data are semi-continuous; and, that mortality is a competing risk. While each of these phenomena have been addressed in the literature, no methods exist for simultaneously acknowledging them. In this work we present a general-purpose modeling framework for nursing home data, with analyses based within the Bayesian paradigm. The framework is illustrated using data from the Long Term Care Focus archive, a multi-state, multi-year, longitudinal database of 20 million nursing home residents integrated with Medicare claims data.

2:10-2:35 PM

Individualized Empirical Null for Profiling Healthcare Providers

Zhi He, Department of Biostatistics, University of Michigan

Existing methods for healthcare provider profiling typically assume that the risk adjustment is perfect and the between-provider variation is entirely due to the quality of care. However, in practice, even with very good models for risk adjustment, there will be characteristics of patients and perhaps providers that are not completely accounted for (e.g.

unobserved socio-economic factors and comorbidities), and many of these characteristics will be related to the outcome and vary across providers. Thus, some of the between-provider variation in a quality measure will typically be due to this incomplete risk adjustment (or unmeasured confounders), which should be recognized in assessing and monitoring providers. Otherwise, conventional methods disproportionately identify larger providers, although they need not be ``extreme''. To fairly assess providers, we propose an individualized empirical null method that accounts for the unexplained variation between providers.

2:35-3:00 PM

Semiparametric Additive Modeling of the Restricted Mean Survival Time

Yuan Zhang, University of Pennsylvania Perelman School of Medicine

Analysis of the restricted mean survival time (RMST) has become increasingly common in biomedical studies of survival. Advantages of RMST over the hazard ratio (HR) include interpretability and lack of reliance on the proportional hazards assumption. Some authors have argued that the RMST should replace hazard regression as the go-to analysis. However, in order for use of the RMST to be more mainstream, it is necessary to broaden the range of data structures to which pertinent methods can be applied. We address this issue from two angles. First, most existing methodological development for directly modeling RMST has focused on multiplicative models. An additive model may be preferred due to goodness of fit and/or parameter interpretation. We propose stratified additive models for direct estimation of RMST. Second, under the proposed methods, categorical adjustment covariates (perhaps high-dimensional) can be factored out of the estimation (akin to stratification in a Cox regression), leaving the focus on the parameters of chief interest. Large- and finite-sample properties are evaluated, and the proposed methods are applied to liver transplant data.

3:00-3:25 PM

Multivariate Spatiotemporal Functional Principal Components Analysis for Modeling Hospitalization and Mortality Rates in the Dialysis Population

Damla Senturk, University of California, Los Angeles

Dialysis patients experience frequent hospitalizations and a higher mortality rate compared to other Medicare populations, in whom hospitalizations are a major contributor to morbidity, mortality, and healthcare costs. Patients also typically remain on dialysis for the duration of their lives or until kidney transplantation. Hence, there is growing interest in studying the spatiotemporal trends in the correlated outcomes of hospitalization and mortality among dialysis patients as a function of time starting from transition to dialysis across the U.S. Utilizing national data from the United States Renal Data System (USRDS), we propose a novel multivariate spatiotemporal functional principal component analysis model (MST-FPCA) to study the joint spatiotemporal patterns of hospitalization and mortality rates among dialysis patients. The proposal is based on a multivariate Karhunen-Loeve expansion that describes leading directions of variation across time and induces spatial correlations among region-specific scores. Novel applications to the USRDS data highlight hot spots across the U.S. with higher hospitalization and/or mortality rates and time periods of elevated risk.

63. INTEGRATIVE ANALYSIS OF MULTI-MODAL NEUROIMAGING AND MULTI-OMICS DATA

Organizer/Chair: Tianzhou Ma, University of Maryland

1:45-2:10 PM

Integrating Imaging and Omics Data for Gene Discovery in Alzheimer's Disease

Li Shen, University of Pennsylvania

Alzheimer’s disease (AD) is a national priority, with 5.8 million Americans affected at an annual cost of $250+ billion and no available cure. Effective strategies are urgently needed to discover new AD genes for disease modeling and drug development. Studying AD genetics using multimodal imaging and multi-omics data is becoming a rapidly growing field with distinct advantages in power over categorical diagnosis under imaging and omics traits as well as in capturing new insights into disease mechanism and heterogeneity from genetic determinants to omics-level molecular signatures, to brain imaging biomarkers, and to AD outcomes. In this talk, we will discuss statistical and informatics strategies for discovering AD risk and protective genes through analyzing multidimensional genetics, omics, imaging and outcome data from landmark and local AD biobanks. We show that the wide availability of these rich biobank data, coupled with advances in biomedical statistics, informatics and computing, provides enormous opportunities to contribute significantly to gene discovery in AD and to impact the development of new diagnostic, therapeutic and preventative approaches.

2:10-2:35 PM

Deep Learning for Feature Extraction and Causal Inference by Integrating Neuroimaging and Genetic Data

Wei Pan, University of Minnesota

Deep learning has been quite successful for classification of image data. In particular, convolutional neural networks have been applied to imaging data to extract low-to-high level features that are related to some disease, e.g. Alzheimer's diseases. However, these features, as possible hidden/latent confounders, may not be biologically related to the disease. On the other hand, there have been renewed interests in and many successful applications of instrumental variables (IV) regression for causal inference in genetics as demonstrated by the recent popularity of Mendelian randomization analyses. Hence, we propose incorporating the use of instrumental variables to improve the chance that the extracted features in deep learning are likely causal to the disease. The proposed method takes advantage of the promises offered by deep learning and IV regression, while integrating imaging and genetic data. We discuss some preliminary results and challenges in our application to the ADNI brain MRI data and AD GWAS data.

2:35-3:00 PM

Network Analysis of Genetic Effects on Brain Connectome for Nicotine Addiction

Shuo Chen, University of Maryland, School of Medicine

Neuropsychiatric disorders are highly heritable traits. In previous genome-wide association studies, the associations between genetic variants and traits of neuropsychiatric disorders have been established. However, the underlying neuropathological pathways remain unclear. We develop a new statistical method to investigate the systematic effects of alleles on brain connectome networks which consequently leads to nicotine addiction. We estimate the network level pathways from genetic variants to dense brain connectome networks and last to a disorder by imposing l0 penalty on both gene-connectome interactions and connectome networks. The results can reveal polygenic loci associated with brain disorders with corresponding specific brain connectome patterns. We apply this method to UK biobank data.

3:00-3:25 PM

From Linear to Deep Collaborative Learning with Applications to Multi-Modal fMRI and Genomics Data Integration

Yu-ping Wang, Tulane University

Canonical correlation analysis (CCA) has been used to find correlations between two or multiple data modalities. However, it is unrelated to phenotypes or disease status. On the other hand, regression models can find the association between a phenotype and multi-modal imaging and genomics data but overlook the cross-modal data correlation. To this end, we first propose a collaborative regression to combine both regression and CCA models. Then, we extend it to the deep learning framework by introducing deep collaborative learning (DCL), which includes deep CCA as a special example. As a result, DCL can better combine complex correlations between multiple data sets in addition to their fitting to phenotypes. Finally, we demonstrate its application to brain development study using integrative fMRI and genomics analysis. We show that DCL outperforms several existing models in predicting populations with different ages and intelligence quotients (IQ). This is a joint work with Dr. Wenxing Hu.

64. STATISTICAL METHODS FOR THE ANALYSIS OF MOBILE HEALTH DATA

Organizer: Lucia Tabacu, Old Dominion University
Chair: Ekaterina Smirnova, Virginia Commonwealth University

1:45-2:10 PM

Predictive Modeling with Weakly Labeled mHealth Data to Personalize Psychotherapy

Samprit Banerjee, Weill Medical College of Cornell University

Smartphones provide an interactive interface that can passively measure various aspects of the user's behavior from device sensors, as well as actively collect self-ratings (e.g. mood, stress etc.) obtained via daily ecological momentary assessment. Taken together with traditional clinical assessments, these measures have the potential to provide unique insight into the treatment trajectories of patients with major depressive disorder undergoing psychotherapeutic treatment. Specifically, patient adherence to psychotherapy sessions is a necessary first step to assess barriers of adherence and personalize future sessions in order to improve adherence and therefore efficacy. There are unique challenges of such predictions due to the noisy nature (missing or under-reporting) of passive and active mHealth data. The nature of missing passive data is unique in the sense that the missed labels are not observed. In this talk, I will introduce these and other challenges of mHealth data analysis and propose semi-supervised machine learning algorithms to address these challenges.

2:10-2:35 PM

Improving the Efficiency of Time-Varying Causal Effect Moderation Analysis in Mobile Health

Walter Dempsey, University of Michigan

Twin revolutions in wearable technologies and smartphone-delivered digital health interventions have significantly expanded the accessibility and uptake of mobile health (mHealth) interventions in multiple domains of health sciences. Sequentially randomized experiments called micro-randomized trials (MRTs) have grown in popularity as a means to empirically evaluate the effectiveness of mHealth intervention components. MRTs have motivated a new class of causal estimands, termed "causal excursion effects", that allow health scientists to answer important scientific questions about how intervention effectiveness may change over time or be moderated by individual characteristics, time-varying context, or past responses. In this talk, we revisit the estimation of causal excursion effects and present two new tools for improving efficiency. Theoretical comparisons accompanied by extensive simulation experiments demonstrate the relative efficiency gains. Practical utility of the proposed methods is demonstrated by analyzing data from a multi-institution cohort of first year medical residents in the United States.

2:35-3:00 PM

Quantile Regression with a Mixture of Function-Valued and a Scalar-Valued Covariate Prone to Classical Measurement Error

Roger Zoh, Indiana University at Bloomington

Current recommendations for dietary intake (DI) and physical activity (PA) to minimize risks for chronic health conditions are based on statistical analyses of data prone to measurement error, including those collected from self-reported questionnaires and wearable devices. Self-reported measures based on food frequency questionnaires are often used in DI assessments, however, they are prone to recall bias. Wearable devices enable the continuous monitoring of PA but generate complex functional data with poorly characterized systematic Tianyerrors. We propose the quantile regression model with function- and scalar- valued covariates prone to measurement errors. We develop semiparametric and parametric approaches to correct for measurement errors associated with the mixture of functional and scalar covariates prone to errors in quantile regression settings. Simulations are performed to assess the finite sample properties of the proposed methods. The developed methods are applied to investigate the influence of wearable-device-based PA and self-reported measures of total caloric intake on quantile function of body mass index (BMI).

3:00-3:25 PM

Mediation Analysis with Quantile Functions as Mediators with an Application to iCOMPARE Trial

Jingru Zhang, University of Pennsylvania

Physical activity has long been shown to be associated with biological and physiological performance and risk of diseases. It is of great interest to assess whether the effect of an exposure or intervention on an outcome mediated through physical activity. However, existing methods for mediation analysis focus almost exclusively on mediation variable that is in the Euclidean space, which cannot be applied directly to the actigraphy data of physical activity. Such data is best summarized in the form of a random histogram or random quantile function. In this paper, we develop a structural equation model (SEM) to the setting where a random quantile function is treated as the mediator to study the indirect mediation effect through physical activity. We provide sufficient conditions for identifying the average causal effects of a quantile function mediator and present methods for estimating the direct and mediating effects with a quantile function being the mediator. We apply our method to the data set from the iCOMPARE trial to explore the mediation effect of physical activity on the causal path between flexible duty-hour policies and sleep related outcomes.

65. NEW STATISTICAL APPROACHES TO COMPLEX MULTI-MODAL BRAIN IMAGING DATA

Organizer: Jun Young Park, University of Toronto
Chair: Danni Tu, University of Pennsylvania Perelman School of Medicine

1:45-2:10 PM

Classification models for multi-modal neuroimaging data with complex geometric structure

Eardi Lila, University of Washington

Multi-modal brain imaging offers a unique opportunity to characterize the structural and functional changes that occur during disease progression. However, when it comes to their statistical analysis, the high dimensionality and complex structure of these data limit the application of standard approaches. We, therefore, introduce a novel classification framework that adopts a Riemannian modeling approach to account for the non-Euclidean structure of the data and is able to constrain the discriminant direction estimates to the space of "physiologically plausible" solutions, ultimately leading to more accurate models. The model proposed is then applied to a pooled dataset from the Alzheimer's Disease Neuroimaging Initiative and the Parkinson's Progression Markers Initiative to identify subjects with Alzheimer's disease from their cortical surface geometry and associated cortical thickness map. We show that our statistical analysis leads to estimated discriminant directions that are interpretable and consistent with the existing neuroscience literature.

2:10-2:35 PM

Spatially-Enhanced Clusterwise Inference for Testing and Localizing Intermodal Correspondence

Jun Young Park, University of Toronto

With the increasing availability of neuroimaging data from multiple modalities-each providing a different lens through which to study brain structure or function-new techniques for comparing, integrating, and interpreting information within and across modalities have emerged. Despite recent methods for testing associations between neuroimaging modalities, they cannot be used to answer questions about where in the brain these associations are most pronounced. In this talk, we introduce a new powerful method that can be used both to test intermodal correspondence throughout the brain and to localize this correspondence. Our method involves first adjusting for the underlying spatial autocorrelation structure within each modality to construct a map of spatially enhanced test statistics. Using structural and functional MRI data from the PNC study, we conduct simulations and data analyses where we illustrate the high statistical power and nominal type I error levels of our method. By constructing an interpretable map of group-level correspondence using spatially-enhanced test statistics, our method offers insights beyond those provided by earlier methods.

2:35-3:00 PM

Continuous and Atlas-free Analysis of Brain Structural Connectivity

Zhengwu Zhang, University of North Carolina at Chapel Hill

Brain structural networks are often represented as discrete adjacency matrices with elements summarizing the connectivity between pairs of regions of interest (ROIs). These ROIs are typically determined apriori using a brain atlas. The choice of atlas is often arbitrary and can lead to a loss of important connectivity information at the sub-ROI level. This work introduces an atlas-free framework that overcomes these issues by modeling brain connectivity using smooth random functions. In particular, we assume that the observed pattern of white matter fiber tract endpoints is driven by a latent random function defined over a product manifold domain. With this representation, we are interested in 1) aligning brain connectomes; 2) deriving connectome-driven brain parcellations, and 3) associating the brain connectomes with human traits. I will introduce our recent development along these three directions.

3:00-3:25 PM

Longitudinal Imaging Data Integration

Seonjoo Lee, New York State Psychiatric Institute and Columbia University

This talk considers longitudinal multimodal fusion via canonical correlation analysis for two longitudinal variables that are possibly sampled at different time resolutions with irregular grids. We modeled trajectories of the multivariate variables using random effects and found the most correlated sets of linear combinations in the latent space. Our numerical simulations showed that the longitudinal canonical correlation analysis (LCCA) effectively recovers underlying correlation patterns between two high-dimensional longitudinal data sets. Finally, we applied the proposed LCCA to the Alzheimer’s Disease Neuroimaging Initiative data and identified the longitudinal profiles of morphological brain changes and amyloid cumulation.

66. IMS MEDALLION LECTURE

Organizer: Xuan Bi, University of Minnesota
Chair: Xihong Lin, Harvard University

1:45-3:30 PM

Statistical Issues in Genome Wide Association Studies

Hongyu Zhao, Yale University

The past two decades have seen great advances in human genetics with the identifications of many genomic regions associated with thousands of diseases through Genome-Wide Association Studies (GWAS) that collect phenotype and genotype data from large cohorts. These data present great opportunities for identifying functional genes and variants for different diseases, inferring relevant tissues and cell types, characterizing genetic architecture, developing risk prediction models, investigating genetic similarities across groups), and studying causal relationships among diseases. There are many challenges in GWAS analysis due to the low signal noise ratios, dependence among genetic markers, complex relationships among traits, the lack of access to individual level data, and the need to incorporate prior knowledge on diseases and pathways, as well as the diverse sources of data generated from international efforts that can facilitate GWAS data analysis. In this presentation, we will discuss methodology developments to address these challenges. The usefulness of the developed statistical methods will be illustrated through their applications to GWAS data on various diseases.

67. CONTRIBUTED PAPERS: COMPUTATIONAL METHODS

Chair: Corinne McGill, The Medical University of South Carolina

1:45-2:00 PM

BayesDIP: An R Package for Bayesian Early Termination in Phase II Clinical Trials Using Decreasingly Informative Priors

Chen Wang, Virginia Commonwealth University

The decreasingly informative prior (DIP) is considered for use in early termination, where null skepticism is explicitly incorporated into the prior in a manner that decreases its prior effective sample size (ESS) as subjects are accrued. The posterior distribution is thus increasingly informed by observed data and less by prior information as a trial continues. This prior formulation restricts adaptation early in a trial, gradually permitting more adaptation as the Bayesian model transfers ESS from the skeptical prior to the likelihood. We show how to parameterize the DIP based on ESS and provide examples. We also present the R package BayesDIP, which provides user-friendly functions for early termination Phase II trial sample size justification using standard Bayesian approach or the DIP approach for binomial, Poisson, and normal distributions. Our functions can help users identify admissible designs (80% power; 5% type I error) for early termination Phase II trials based on determining planned total sample size, expected sample size and its variance, and efficacy and futility boundaries, and can also estimate power and type I error for any set of specifications.

2:00-2:15 PM

Exact MCMC Free Bayesian Inference for a Class of Spatial Generalized Linear Mixed Effects Models

Jonathan Bradley, Florida State University

Markov chain Monte Carlo (MCMC) has become a standard in Bayesian statistics that allows one to generate dependent replicates from a posterior distribution for general Bayesian hierarchical models. However, convergence issues, tuning, and the effective sample size of the MCMC are nontrivial considerations that are often overlooked or can be difficult to assess. This motivates us to consider finding expressions of the posterior distribution that are computationally straightforward to sample from directly (i.e., independently) without MCMC. We focus on a broad class of Bayesian generalized linear mixed-effects models (GLMM) that allows one to jointly model data of different types. We derive a class of distributions that allows one to specify the prior on fixed and random effects to be any conjugate multivariate distribution. The expression of the posterior distribution is given, and direct simulations have an efficient projection form. An analysis of an environmental spatial dataset is presented.

2:15-2:30 PM

Using Deep Transfer Learning to Identify Markers of the Epithelial-to-Mesenchymal Transition in Pancreatic and Prostate Cancer Spatial Transcriptomics

Justin Couetil, Indiana University School of Medicine

DEGAS identifies patterns in patient transcriptomic and clinical data and maps these associations on to higher-resolution data like spatial and single-cell transcriptomics. We provide a baseline comparison with a similar tool. Then, we the apply DEGAS to prostate and pancreatic tissues, identifying transcriptomic signatures enriched for ontology terms associated with growth regulation and apoptosis, inflammation, immune signaling, and autophagy in histologically normal prostate tissues and adjacent normal pancreatic cancer tissue. The regions highlighted by DEGAS could reflect transcriptional precursors to intraepithelial neoplasia; a well-recognized premalignant morphological change in tissues. Identifying biomarkers of tissue stress that precede morphologic diagnosis of high-grade pre-malignant lesions by a pathologist may help triage patients at high risk for future development of cancer, or to better understand whether histologically normal pre-malignant tissues at tumor margins contribute to recurrence. DEGAS is a key tool for hypothesis generation from single cells and spatial transcriptomics, which are information-rich assays with small sample sizes.

2:30-2:45 PM

Bioconductor Infrastructure for Analyzing Multiplex Single Cell Imaging Data in R

Julia Wrobel, Colorado School of Public Health

We introduce a new software ecosystem in R for data storage and spatial analysis of multiplex single-cell tissue imaging data. Our central R package is called spatialMI and builds on SpatialExperiment, which is an R package and S4 class on Bioconductor that provides a special data infrastructure for spatially resolved transcriptomics data that facilitates data storage, retrieval, subsetting, and interfacing with downstream tools. The spatialMI package adapts data structures from the SpatialExperiment class and inherits methods from that and the popular SingleCellExperiment class so that spatialMI users can easily access software developed for other similar single cell data types. For multiplex single cell imaging data specifically, we build additional S4 methods to convert multichannel tiff images to a novel spatialMI data class so data from any multiplex platform, including CODEX, Vectra-Polaris, MIBI, IMC, etc, can be converted into the same type of Bioconductor data object. This facilitates consistency and ease of use in downstream analysis. We discuss the available functionality as well as forthcoming extensions in the form of satellite packages.

2:45-3:00 PM

Robust Effect Size Index Analysis Using RESI R Package

Megan Jones, Vanderbilt University

Effect size indices are useful parameters that quantify the strength of association and are unaffected by sample size. There are many available effect size parameters and estimators, but it is difficult to compare effect sizes across studies as most are defined for a specific population parameter. We introduced a new effect size measure, the Robust Effect Size Index (RESI), which is advantageous because it is defined from M-estimators, so is not model-specific, uses a robust covariance estimate and can be used in a wide range of models. We recently developed confidence interval procedures for the RESI, useful for quantifying the estimate's precision. We present the RESI R package, which makes it easy to report the RESI and its confidence interval alongside common analyses. The package produces coefficient and ANOVA tables and overall Wald tests for many model types, appending the RESI estimate and confidence interval to each. The package also includes functions for conversion, visualization and interpretation. We will briefly introduce the RESI estimators and confidence interval procedure and walk through practical application with a demonstration of the RESI R package.

3:00-3:15 PM

SPARTIN: A Bayesian Method for the Quantification and Characterization of Immune Cell Infiltration in Spatial Pathology Data

Nathaniel Osher, University of Michigan

The large and ever increasing volume of readily available pathological imaging data has sparked the development of methods to extract features from such data. In particular, the rise of such data has led to an increased exploration of the tumor microenvironment with respect to the presence and spatial composition of immune cells. Spatial statistical modeling of the immune microenvironment may yield insights into the role played by the immune system in the natural development of cancer as well as treatment opportunities for patients. In this paper, we present the SPatial Analysis of paRtitioned Tumor-Immune imagiNg (SPARTIN) method, a Bayesian method for the quantification of immune cell infiltration. SPARTIN uses Bayesian Strauss processes to characterize a novel measure of tumor-immune cell interaction within and across biopsies. We applied the method to data from 335 Skin Cutaneous Melanoma images from the Cancer Genome Atlas and found that at the biopsy level tumor-immune cell spatial association was inversely associated with survival and pathologist assessment of lymphocyte presence. In addition, we found significant associations with genomic and clinical outcomes.

3:15-3:30 PM

Cell Composition Inference and Identification of Layer-Specific Transcriptional Profiles with POLARIS

Jiawen Chen, University of North Carolina at Chapel Hill

Spatial transcriptomics (ST) technology, providing spatially resolved transcriptional profiles, facilitates advanced understanding of key biological processes related to health and disease. Sequencing-based ST technologies provide whole-transcriptome profiles, but are limited by the non-single cell level resolution. Lack of knowledge in the number of cells or cell type composition at each spot can lead to invalid downstream analysis, which is a critical issue recognized in ST data analysis. Methods developed, however, tend to under-utilize histological images, which conceptually provide important and complementary information including anatomical structure and distribution of cells. To fill in the gaps, we present POLARIS, a versatile ST analysis method that can perform cell type deconvolution, identify anatomical or functional layer-wise differentially expressed (LDE) genes and enable cell composition inference from histology images. Applied to four tissues, POLARIS demonstrates high deconvolution accuracy, accurately predicts cell composition solely from images, and identifies LDE genes that are biologically relevant and meaningful.

68. CONTRIBUTED PAPERS: SURVIVAL ANALYSIS

Chair: Varadan Sevilimedu, Georgian Southern University

1:45-2:00 PM

Comparison of Statistical Methods for the Case and Extreme-Control Design of Time-to-Event Data

Wenan Chen, St. Jude Children's Research Hospital

It remains expensive to apply next generation sequencing technology to a large number of in genomic association studies. For time-to-event phenotypes, two classical cost effective designs are nested case-control design and case-cohort design, which sequence all cases and only a subset of controls. Recently, it has been shown that instead of randomly sampling controls as in the two classical designs above, sampling extreme controls may increase the statistical power. Here we study a specific case versus extreme-control design, which includes all cases and extreme controls. Each extreme control is a sample without the event and the censoring time is above a certain threshold. Through simulations, we compare the power, type I error among different analysis methods for the case and extreme-control design, including logistic regression and cox regression. We also compare the power between the case and extreme control design and the nested case control design. Finally, we show its application in a retrospective case and extreme-control study to identify the associations between genomic somatic alterations at diagnosis and later relapse of pediatric acute lymphoblastic leukemia.

2:00-2:15 PM

Estimating the Proportional Hazards Cure Model with Background Mortality on Interval Censored Data

Shujie Chen, University of South Carolina

With the development of treatment, patients may be cured and suffer from other cause of death. The cure model with background mortality can measure the population cure which refers to the patients with comparable mortality with their counterpart in general population. In this paper, we extend the background mortality proportional hazards (PH) cure model to the interval censoring data, which can be applied to the case when the time to death is recorded imprecisely. For the estimation, we develop the EM algorithm with monotone spline function, a flexible parameter method. The proposed methods were evaluated via the comprehensive simulation studies and further being applied to SEER female breast cancer data set.

2:15-2:30 PM

Causal Inference for Survival Outcome with Functional Treatment

Xiyuan Gao, University of Missouri, Columbia

We consider a functional causal accelerated failure time(AFT) model for discovering causality between functional predictors and time-to-event outcomes. The propensity score for functional treatments is properly defined in terms of a multivariate substitute under both parametric and nonparametric frameworks. The weighted estimator, regression adjustment estimator, and double robust estimator are proposed based on the estimated functional propensity score. The appealing performance of the proposed method is demonstrated by a simulation study. The proposed method is applied to study the causal effect of magnetic resonance imaging(MRI) on survival outcomes in Alzheimer?s Disease.

2:30-2:45 PM

Apply Deep Learning Method in Mixture Cure Rate Model

Xiaowen Sun, University of South Carolina

The mixture cure model (MCM) becomes more popular in handling disease with potential cured. The two components MCM includes the incidence and latency components which are modeled by the logistic and Cox proportional hazards regression respectively. The application of mixture cure model to Electronic Health Records (EHRs) data might be challenge due to the high dimensional, non-linearity, and huge volume of EHRs. We propose to incorporate the advanced machine learning method (fully connected layer and convolutional neural network) in MCM estimation which accounts for non-linear association of predictors with high dimension. The Expectation-Maximization algorithm is used to estimate the parameter of interest in MCM. Using this approach, we assessed the model performance by mean of area under curve (AUC) of the non-cured probability, mean square error (MSE) of baseline survival, MSE of non-parametric predictors in latency part for test set. Simulation studies show good performance in both deep learning method than linear method. We will apply the COVID-19 vaccine data for real data analysis.

2:45-3:00 PM

A Flexible Copula Model for Dependent Bivariate Survival Data with Informative Censoring

Reuben Adatorwovor, University of Kentucky

Censoring in time-to-event survival analysis is a common way of incorporating available data in the analysis of right-censored survival data. The existing literature assumes that the exact event time and the censoring time are independent and non-informative. An assumption that may be unsubstantiated and generally introduce biases in the estimation procedure. Popular methods like Inverse Probability Weighting (IPW) and other variants attempt to correct the bias associated with the independent censoring assumption in a univariate setting. We proposed a copula-based dependent censoring methodology for bivariate time-to-event data, which models the event and censoring times through a copula-based Cox proportional hazard model formulation. Our estimator possesses strong consistency and desirable asymptotic properties under regularity conditions. We provided results under extensive simulations with application to the Danish twin prostate cancer data set.

3:00-3:15 PM

Inference for Relative Hazard and Covariate-specific Pure Risk Estimated from Stratified and Unstratified Case-Cohort Data

Lola Etievant, Biostatistics Branch, DCEG, NCI, NIH

Case-cohort sampling permits estimation of relative hazard (RH) and pure risk (PR) under the Cox model. It allows analysis of multiple endpoints and reduces costs as covariates need only be measured for subjects with an event (cases) and for non-cases in a random sample from the cohort (subcohort). Selecting the subcohort by stratified random sampling and weighting sampled non-cases by the inverse of stratum-specific sampling fractions increase efficiency of RH (Borgan et al., 2000). Calibrating weights with information available on the whole cohort can further improve efficiency (Breslow et al., 2009). While stratified sampling requires intricate variance estimation, many researchers use the simpler "robust variance" (V1) proposed for unstratified designs (Barlow, 1994). We investigated how V1 performs for RH and PR in the various designs, compared to the theoretically correct estimator, V2 (Samuelsen et al., 2007). We provided flexible influence-based variance estimates for RH and PR with V1 and V2. Although V1 works in many situations, we recommend V2 that properly accounts for the sampling features.

3:15-3:30 PM

Two-Phase Outcome-Auxiliary-Dependent Sampling with Failure Time Data

Xu Cao, University of North Carolina at Charlotte

Epidemiological studies often seek to relate time to a failure event to some exposure variables that are expensive to obtain, thus large cohort studies under simple random sampling could be prohibitive to conduct with limited budget. Outcome-dependent sampling (ODS) is a commonly used cost-effective sampling strategy in such studies. To further enhance study efficiency upon ODS, we propose a two-phase outcome-auxiliary-dependent sampling (OADS) design by incorporating cheaply available auxiliary variables. It allows the probability of obtaining the expensive exposures to depend on both the failure time and auxiliary variables. To account for the sampling bias, we develop a two-step pseudo-likelihood approach for inference and a non-parametric bootstrap procedure for variance estimation. The proposed method is shown to be more efficient than other competing sampling schemes. Its application to an epidemiological study is provided.

69. CONTRIBUTED PAPERS: ADVANCES IN CAUSAL INFERENCE

Chair: Giulio Grossi

1:45-2:00 PM

Multi-Study R-Learner for Heterogeneous Treatment Effect Estimation

Cathy Shyr, Vanderbilt University Medical Center

Flexible estimation of heterogeneous treatment effects is central to precision medicine. While efforts in systematic data sharing and data curation initiatives have increased access to multiple datasets, existing methods for estimating heterogeneous treatment effects are largely rooted in theory based on a single study. In this work, we extend the R-learner to the multi-study setting and propose a general class of two-step algorithms for treatment effect estimation. In particular, the multi-study R-learner generalizes the R-learner to achieve cross-study robustness of confounding adjustment. This approach is not only easy to implement but also highly flexible due to 1) the adoption of series estimators for the treatment effects and 2) the incorporation of modern machine learning techniques for the estimation of nuisance functions. We show the estimator is asymptotically normal. Moreover, we illustrate via simulations and a breast cancer data application that the multi-study R-learner results in lower estimation error than the R-learner as between-study heterogeneity in confounding adjustment increases.

2:00-2:15 PM

Impact of Missing Data and Imputation Methods in Causal Structural Learning

Jiarui Lu, Novartis Pharmaceuticals Corporation

Statistical methods for causal structural learning are often used to undercover the underlying directed acyclic graphs (DAG) from observational dataset. PC-algorithm and greedy equivalence search (GES) are two well-known methods for causal structural learning. Both are built up for causal discovery from complete observational datasets. One study has identified the existence of bias for causal discovery when the missing of a variable depends on the collider. This suggests an connection between the bias of causal discovery and the patterns of missing data, but the performance difference of current methods using data with different missing patterns is still unknown. We conducted a comprehensive study to evaluate the performance of both PC-algorithm and GES method with different missing mechanisms including missing complete at random and missing at random. Our simulations suggest a bias of both methods using complete case data or imputations methods for causal structural learning, when the missing of one variable is depends on its parents and child. Also, the GES method works better than PC algorithm when the underlying DAG is relatively sparse.

2:15-2:30 PM

Outlier Resistant Inference for Conditional Average Treatment Effect

Ran Mo, IUPUI

The estimation of causal effect based on conditional average treatment effect (CATE) is usually vulnerable to outliers. However, to the best of our knowledge, the outlier-resistant inference for the CATE has not been investigated in the literature. In this work, we propose an outlier-resistant estimation method for the CATE by incorporating M-estimation in the inverse propensity weighting (IPW) approach. The influence function and breakdown property are investigated to study the robustness of our method. In addition, we derived the asymptotic properties of the proposed estimator for inference purposes. The finite sample performance of the proposed estimator is evaluated via Monte Carlo experiments. The proposed method is compared with the IPW method and the augmented inverse probability weighting (AIPW) method, which do not account for outliers. Finally, the proposed method is applied to the NCSCHS dataset and the NHANES dataset to estimate the average effects of smoking on birth weights and the White blood cell count condition on age, respectively.

2:30-2:45 PM

Optimizing Event-Triggered Adaptive Interventions in Mobile Health with Sequentially Randomized Trials

Mason Ferlic, Data Science for Dynamic Decision-making (D3C), University of Michigan

In mobile and digital health, advances in collecting sensor data and engaging users in self-reporting have made it possible to monitor an individual?s response to intervention in real-time. This has led to an interest in event-triggered adaptive interventions, in which a patient transitions to the next stage of treatment when pre-specified event criteria are triggered. Sequential, multiple-assignment randomized trial (SMART) designs can be used to develop optimized event-triggered adaptive interventions. We introduce a new estimation approach for analyzing data from SMARTs which addresses four statistical challenges: (i) the need to condition on the event trigger, (ii) while avoiding causal collider bias in the comparison of adaptive interventions starting with different treatments; and the need for dimension-reducing models for (iii) the distribution of the event given the past and (iv) the relationship between the event and the research outcome, all while avoiding negative impacts of model misspecification bias on the target causal effects. The method is illustrated using data from a SMART to develop an event-triggered adaptive intervention for weight loss.

2:45-3:00 PM

An Information-Theoretic Approach for the Assessment of a Continuous Outcome as a Surrogate for a Binary True Endpoint Based on Causal Inference: Application to Vaccine Evaluation

Fenny Ong, Hasselt University

The development of methods to validate surrogate endpoints remains an active field of research due to its importance on the impact of accelerating the approval of a large number of new promising treatments. Within the causal association paradigm, we propose a method to assess the validity of a continuous outcome as a surrogate for a binary true endpoint. Based on causal-inference concepts, a new model is proposed to describe the joint distribution of the potential outcomes of the putative surrogate and the true endpoint of interest. The identifiability issues inherent to this type of model are handled via sensitivity analysis. Subsequently, using the information-theoretic ideas, a metric of surrogacy, the so-called Individual Causal Association (ICA), is presented. We argue that the proposed metric has convenient theoretical properties and a simple yet intuitive interpretation. Further evaluation of the methodology via simulations and its implementation in a randomized clinical trial evaluating an inactivated quadrivalent influenza vaccine also showed reasonable results.

3:00-3:15 PM

Causality Inference in EEG Data with Regime Switching Time Series Modeling Approach

Sipan Aslan, King Abdullah University of Science and Technology (KAUST), Statistics, CEMSE Division

Exploring and revealing the underlying complex interactions between brain regions during neuronal activities is an active research area in neuroscience. There is a need for sophisticated statistical models to advance research on the brain's highly complex information processing system. In this direction, many studies in the literature deal with causality and connectivity in EEG time series from many perspectives and purposes. There stands an essential need for statistical methods that provide interpretable outputs, as well as the procedures that include the Black-Box routines, which have mainly been developed as the volume of data to be processed has increased in recent years. In light of many studies that have been done so far, we aimed to implement Threshold Autoregressive (TAR) modeling for nonlinear causality inference by analyzing the publicly available EEG dataset at the PhysioNet platform observed during the motor movement/imagery experiment. In this study, in addition to presenting exploratory analyses of the dataset, we demonstrate the advantages of regime-switching TAR modeling as a promising method to investigate nonlinear causality.

3:15-3:30 PM

Nonparametric Doubly Robust Inference for Average Treatment Effect on the Treated with Missing Outcomes

Lindsey Schader, Emory University

In many policy and medical applications, there is interest in estimating the average treatment effect among the treated (ATT). Existing estimators of the ATT generally rely on the assumption that two regression quantities converge relatively quickly to their respective true values. This assumption underlies the validity of confidence intervals and hypothesis tests derived from these estimators. In this paper we propose a targeted maximum likelihood estimator (TMLE) for the ATT that provides two major contributions: (i) the estimator accounts for missing at random outcomes; (ii) the estimator enjoys a double robust asymptotic limiting distribution under weaker convergence assumptions than those typically required. These weaker assumptions are particularly beneficial in settings where flexible regression techniques, such as machine learning, are used in the estimation process. We demonstrate the properties of this estimator with a simulation study.

70. CONTRIBUTED PAPERS: MULTIVARIATE/NONLINEAR MODELS AND METHODS

Chair: Yaomin Xu, Vanderbilt University

1:45-2:00 PM

A Mixed Effects Bayesian Regression Model for Multivariate Group Testing Data

Christopher McMahan, Clemson University

Laboratories use group (pooled) testing with multiplex assays as a means to reduce the time and cost associated with screening large populations for infectious diseases. Group testing reduces cost by testing pooled specimens for the presence of an infectious agent. When combined with multiplex assays, which screen for multiple diseases simultaneously, group testing offers a more timely and cost effective testing protocol, when compared to traditional implementations. These benefits come at the expense of a more complex data structure, which could hinder surveillance efforts. To overcome this challenge, we develop a general Bayesian methodology that can be used to fit a mixed multivariate probit model to data arising from any group testing protocol that makes use of a multiplex assay. In the formulation of this model, we account for the correlation between the disease statuses, the heterogeneity across population subgroups, and provide for automated variable selection through the adoption of spike and slab priors. To complete model fitting, we develop an easy to implement posterior sampling algorithm.

2:00-2:15 PM

An Adaptive CUSUM Chart for Drift Detection

Fan Yi, University of Florida

In practice, sequential processes often have gradual changes in their process distributions over time. This is related to the drift detection problem in statistical process control. In the literature, there have been some existing discussions on this problem. But, most existing methods are designed based on the assumption that the related drift is linear or have another specific pattern. In reality, however, such specified patterns may not be valid. In this paper, we suggest an adaptive CUSUM chart to handle the drift detection problem with a flexible drift pattern. The new method integrates the general framework to construct a CUSUM chart based on the generalized likelihood ratio statistic and estimation of a shift size by the exponentially weighted least square regression procedure. Simulation studies show that the proposed method is effective in various cases considered. The new method is also illustrated using an example about the exchange rates between Indian Rupees and US Dollars.

2:15-2:30 PM

Matrix Linear Models for Connecting Lipid Composition to Individual Characteristics

Gregory Farage, University of Tennessee Health Science Center

We have developed Matrix Linear Models (MLM), a family of bilinear models, for studying associations in structured high-throughput data. The method extracts information by aggregating signals through the rows (samples) of the matrix data and across the columns (omics features) in a single model. It contrasts with the standard approach in lipidomics, where each lipid is analyzed singly for associations with sample features, and a second analysis is done to look for patterns among features showing common associations. We demonstrate how MLM offers flexibility, computational speed, and the power to detect associations by applying our method to three lipidomics studies. The MLM framework can estimate relationships in lipids sharing known characteristics, whether categorical (e.g., type of lipid or pathway) or numerical (e.g., the number of double bonds in triglycerides). We show how our method can separate the contributions of two correlated triglyceride features: the number of carbon atoms and double bonds that would be missed if we analyzed lipids individually. Our method was implemented using the Julia package, MatrixLM (https://github.com/senresearch/MatrixLM.jl).

2:30-2:45 PM

Modeling Disease Transition Based on EHR Data: A Bayesian Change Point Analysis

Yunju Im, University of Nebraska Medical Center

Diseases can be interconnected, and the occurrence of one disease can lead to the occurrence of itself and other diseases in the future. Research has been conducted on the “transition” of diseases over time. In this study, we consider a special temporal transition structure, under which there are change points; between two adjacent change points, the transition model remains unchanged; and the transition models on the two sides of a change point differ. It is further recognized that the transition models are sparse. To achieve simultaneous estimation/identification of sparsity as well as clustering (identification of change points), a Bayesian approach is proposed. The analysis of the Taiwan NHI (National Health Insurance) data leads to biomedically sensible findings.

2:45-3:00 PM

Meta-Analysis for Modeling Studies with Multiple Cut-Points and a Simulation Study

Feng Zhang, The University of Texas MD Anderson Cancer Center

In medicine, it is important to identify the predictive effect of a biomarker on treatment to select the best treatment. For example, PD-L1 has been shown to predict the success of immunotherapy. Patients with high PD-L1 expression are more likely to respond to checkpoint inhibitors. Our goal is to accurately estimate the individual-level Emax dose-response model of a continuous marker and the outcome using aggregated data reported in papers, e.g., the outcome of PD-L1 low- and high-groups. However, with only aggregated data, the standard method (GLS) yields a biased estimate of the slope parameter describing the steepness of the curve. To reduce the bias, we proposed a one-stage data augmentation method. Three out of four parameters describing minimal/maximal/ED50 can be estimated using MLE, The slope parameter can be further estimated through the proposed method. The proposed method is robust for providing an accurate estimate of the slope parameter in a variety of design settings. Via simulations, we also find that the number of trials and the number/location of cut-points can impact the accuracy of estimation. Results of the analysis on real data will also be reported.

3:00-3:15 PM

Prolong: Penalized Regression on Longitudinal Omics Data with Network and Group Lasso Constraints

Steve Broll, Cornell University

There is a growing interest in longitudinal omics data, but there are gaps in existing methodology in the high-dimensional setting. This presentation focuses on selecting metabolites that co-vary with Tuberculosis viral load. The proposed method is applied to general continuous longitudinal phenotypes with longitudinal omics predictors. Simple longitudinal models examining a single omic predictor at a time do not leverage the correlation across predictors, thus losing power. We propose a penalized regression approach on the first differences of the data that extends the lasso + Laplacian method (Li and Li 2008) to a longitudinal group lasso + Laplacian approach. Our method, PROLONG, leverages the first differences of the data to address the piecewise linear structure and the observed time dependence. The Laplacian network constraint incorporates the correlation structure of the predictors, and the group lasso constraint induces sparsity while grouping metabolites across their first differenced observations.

3:15-3:30 PM

Multivariate Single Index Modeling of Longitudinal Data with Multiple Responses

Zibo Tian, University of Florida

In medical studies, composite indices are routinely used for predicting medical conditions of patients. These indices are usually developed from observed data of certain disease risk factors, and it has been demonstrated in the literature that single index models can provide a powerful tool for this purpose. In practice, the observed data of disease risk factors are often longitudinal in the sense that they are collected at multiple time points for individual patients, and there are often multiple aspects of a patient?s medical condition that are of our concern. However, most existing single index models are developed for cases with independent data and a single response variable, which are inappropriate for the problem just described in which within-subject observations are usually correlated and there are multiple mutually correlated response variables involved. This paper aims to fill this methodological gap by developing a single index model for analyzing longitudinal data with multiple responses. Both theoretical and numerical justifications show that the proposed new method provides an effective solution to the related research problem.

71. CONTRIBUTED PAPERS: INTEGRATED-OMIC DATA ANALYSIS, SINGLE-CELL/MICRORNA DATA METHODS

Chair: Kate Shutta, Harvard T.H. Chan School of Public Health

1:45-2:00 PM

A Unified Quantile Framework Reveals Nonlinear Heterogeneous Transcriptome-Wide Associations

Tianying Wang, Center for Statistical Science

Transcriptome-wide association studies (TWAS) are powerful tools for identifying putative causal genes by integrating genome-wide association studies and gene expression data. Most existing methods are based on linear models and, therefore, may miss or underestimate nonlinear associations. In this article, we propose a robust, quantile-based, unified framework to investigate nonlinear transcriptome-wide associations in a quantile process manner. Through extensive simulations and the analysis of multiple psychiatric and neurodegenerative disorders, we showed that the proposed framework gains substantial power over conventional approaches and leads to insightful discoveries on nonlinear associations between gene expression levels and traits, thereby providing a complementary approach to existing literature. In doing so, we applied the proposed method for 797 continuous traits from the UK Biobank, and the results are available in a public repository.

2:00-2:15 PM

Bayesian Simultaneous Factorization and Prediction Using Multi-Omic Data

Sarah Samorodnitsky, University of Minnesota Division of Biostatistics

Integrative factorization methods for multi-omic data decompose variation across omics sources. However, most do not quantify uncertainty in the estimated variation structure, nor simultaneously account for a phenotype. We propose two Bayesian approaches to simultaneously factorize variation and perform prediction in a complete framework for uncertainty. We use conjugate normal priors and show that the posterior mode of this model can be estimated by solving a structured nuclear norm-penalized objective that also achieves rank selection (i.e., the dimension of the latent variation structure) and motivates the choice of hyperparameters. BSFP accommodates concurrent imputation and full posterior inference for missing data, including ?blockwise? missingness, and prediction of unobserved outcomes. We show via simulation that BSFP propagates uncertainty from the estimated factorization to prediction, yielding appropriate coverage. We use BSFP to predict lung function based on the metabolome and proteome in bronchoalveolar lavage, revealing a cluster of patients with chronic obstructive pulmonary disease driven by shared proteomic and metabolomic expression.

2:15-2:30 PM

Benchmarking Highly Variable Feature Selection in Single-Cell RNAseq and Beyond

Ruzhang Zhao, Johns Hopkins University

Highly variable feature selection aims at selecting features with high variations after mean-variance adjustment, which is an important and necessary step in the standard analysis pipelines of single-cell RNA sequencing data. Various methods have been proposed in recent years, which are based on different data formats and use different ways of adjusting for the mean-variance relationship. However, the current literature about highly variable feature selection comparison is limited due to the method list, datasets, and evaluation criteria. Here, we comprehensively benchmark highly variable feature selection m?ethods with updated method list, strict evaluation criteria, and sufficient number of updated datasets. Furthermore, based on benchmark results, we propose a new highly variable feature selection method via combining multiple methods, which achieves the best performance. The functions are compiled in R package mixhvg.

2:30-2:45 PM

Gaussian Graphical Model-based Hierarchical Cancer Heterogeneity Analysis via Integrating Pathological Imaging and Omics Data

Mingyang Ren*, Department of Statistics, The Chinese University of Hong Kong

Cancer is heterogeneous in nature. In addition to commonly considered simple data characteristics, recent studies have shown that incorporating interconnections among variables can lead to more informative heterogeneity structures. To this end, Gaussian Graphical Model (GGM)-based approaches have been developed. In most of the existing cancer heterogeneity analysis, only a single type of data is considered. Advancing from the literature, we propose integrating pathological imaging data and omics data. In clinical practice and research, pathological imaging features are usually examined in the first step to provide a rough subgrouping. Omics data can then be analyzed to provide a refined sub-subgrouping. As such, we develop a novel penalization approach and propose reinforcing a hierarchy in heterogeneity analysis, with the sub-subgroups characterized by omics measurements nested in the subgroups characterized by pathological imaging features. Consistency properties are established. In the analysis of TCGA data on lung cancer, clinically meaningful (sub-)subgroups different from the alternatives are identified.

2:45-3:00 PM

Joint Tensor Analysis of Single Cell 3D Genome and Epigenetic Data with Muscle

Kwangmoon Park*, University of Wisconsin-Madison

Emerging single cell technologies that simultaneously capture long-range interactions of genomic loci together with their DNA methylation genomewide are advancing our understanding of three-dimensional (3D) genome structure and its interplay with the epigenome at the single cell level. However, methods that can jointly analyze multiple modalities with single cell high throughput chromatin conformation capture (scHi-C) data are lacking. Here, we introduce Muscle, a semi-non negative joint decomposition of multiple single cell tensors, to jointly analyze single cell 3D confirmation and DNA methylation. Muscle takes advantage of the inherent tensor structure of the scHi-C data and integrates it with genomewide DNA methylation. Parameters estimated by Muscle directly align with the key parameters of the downstream analysis of scHi-C data in a cell type specific way. Evaluations with data-driven experiments demonstrate the advantages of the joint modeling framework of Muscle over single modality modeling for cell type delineation and elucidating associations between modalities.

3:00-3:15 PM

Multiple Augmented Reduced Rank Regression for Pan-Cancer Analysis

Jiuzhou Wang, University of Minnesota

Statistical approaches that successfully combine multiple datasets are more powerful, efficient, and scientifically informative than separate analyses. To address variation architectures comprehensively for high-dimensional data across multiple samples (i.e., cohorts), we propose multiple augmented reduced rank regression (maRRR), a flexible matrix regression and factorization method to concurrently learn both covariate-driven and auxiliary structured variation. We consider a structured nuclear norm objective motivated by random matrix theory, in which the regression or factorization terms may be shared or specific to any number of cohorts. Our framework subsumes several existing methods. Simulations demonstrate substantial gains in power from combining multiple datasets, and from parsimoniously accounting for all structured variation. We apply maRRR to gene expression data from multiple cancer types (i.e., pan-cancer) from TCGA, with somatic mutations as covariates. The method performs well with respect to prediction and imputation of held-out data and provides new insights into mutation-driven and auxiliary variation that is shared or specific to certain cancer types.

72. CONTRIBUTED PAPERS: MEDICAL/WEARABLE DEVICES, AGREEMENT, ROC ANALYSIS

Chair: Jing Kersey, Georgia Southern University

1:45-2:00 PM

Assessing Agreement of Functional Data Among Multiple Methods and Replicates

Jeong Hoon Jang, Yonsei University

In this work, we introduce a series of concordance correlation coefficient (CCC) indices for evaluating the reliability and reproducibility of modern medical devices that produce functional data (e.g., curve or image), whose sampling unit is a smooth continuous function defined over a time or spatial domain. Specifically, intra-CCC index quantifies the agreement among replicates of functional data produced by the same method. Inter-CCC measures the agreement among different methods based on the average of the replicates of functional data. Total-CCC represents the agreement among different methods based on individual functional data. The estimation is based on the multivariate multilevel functional model that expresses observations as method-specific and replicate-specific multivariate functional principal components.

Extensive simulation studies are performed to assess the finite-sample properties of the estimators. The proposed method is applied to Emory renal study data to evaluate the reproducibility and reliability of renogram curves produced by high-tech radionuclide image scans that are used to non-invasively detect kidney obstruction.

2:00-2:15 PM

Semiparametric Models of Inter- and Intra-Individual Variability Through Generalized Distances of Wearable Data

Jinyuan Liu, Vanderbilt University

Longitudinal observations often display large variability, although the linear mixed effects model (LME) designates between- and within-subject variances, the assumed homoscedastic variance across subjects over time rarely holds for wearable measurements. Fitting LME for such data usually yields huge variances, signaling the need to parse them further using covariates. A mixed-effects location scale model (LSM) has been proposed accordingly. However, it not only posits stringent distribution assumptions for the latent random effects that are not validatable but suffers from an extreme computational burden. We hence prescribe a robust and efficient alternative. By leveraging distance metrics to capture variability components, we propose a distance-based regression to discern inter- and intra-individual variability and identify subject- and time-varying risk factors for each. This semiparametric framework retains the ease of parameter interpretation from LSM but gains protection from distribution misspecification. It facilitates computation by leveraging efficient estimating equations for sensitive signal detection, revealing its potential to scale up in mHealth and big data.

2:15-2:30 PM

Bayesian Spatial Cluster Signal Learning for Adverse Events (AE) Detection

Hou-Cheng Yang, FDA/CDRH

There is growing interest in understanding geographic patterns of medical device-related adverse events (AEs). Hu et al. (2021) used a spatial scan method combined with likelihood ratio test (LRT) for spatial-cluster signal detection over the geographical region. They used a moving window to scan the entire study region and collected a bunch of candidate sub-regions from which the spatial-cluster signal(s) will be found. However, it has computational cost challenge. The computational cost may increase if a large spatial-cluster pattern is present, or the number of sub-regions increases. To tackle the computational cost issue, we used a Bayesian nonparametric method that combines the ideas of Markov random field as an alternative approach to leverage geographical information and find potential clusters. Then, we applied the LRT for signal detection on this potential clusters. Furthermore, our method can provide an ability to capture both locally spatially contiguous clusters and globally discontiguous clusters. We not only provide extensive simulation studies but also used the Left Ventricular Assist Device (LVAD) data as an illustration of the effectiveness of this method.

2:30-2:45 PM

Smooth Estimator for the Length of the Receiver Operating Characteristic Curve

Pablo Martinez-Camblor, Geisel School of Medicine at Dartmouth, Dartmouth College

A good diagnostic test should show different behavior on both the positive and the negative populations. However, this is not enough for having good classification systems. The binary classification problem is a complex task, which implies to define decision criteria. The knowledge of the level of dissimilarity between the two involved distributions is not enough for developing a good classification system. We have to know how to define those decision criteria. The length of the ROC curve has been proposed as an index of the optimal discriminatory capacity of a biomarker. It is related not with the actual but with the optimal classification capacity of the considered diagnostic test. One particularity of this index is that its estimation should be based on parametric or smoothed models. We explore here the behavior of a kernel density estimator-based estimator for approximating the length of the ROC curve.

2:45-3:00 PM

Estimation and Inference on the Partial Volume Under the ROC Surface with Applications to Pancreatic Cancer

Kate Young, University of Kansas Medical Center

Summary measures of biomarker accuracy that employ the receiver operating characteristic (ROC) surface have been proposed for biomarkers that classify patients into one of three groups: healthy, benign, or aggressive disease. The volume under the ROC surface (VUS) summarizes the overall discriminatory ability of a biomarker in such configurations, but includes cutoffs associated with clinically irrelevant true classification rates (TCR's). Due to the lethal nature of pancreatic cancer, cutoffs associated with a low TCR for identifying patients with pancreatic cancer may be undesirable and not appropriate for use in a clinical setting. In this project, we study the properties of a more focused criterion, the partial VUS (pVUS), that summarizes the diagnostic accuracy of a marker in the three-class setting for regions restricted to only those of clinical interest. We propose methods for estimation and inference on the pVUS under parametric and non-parametric frameworks and apply these methods to the evaluation of potential biomarkers for the diagnosis of pancreatic cancer.

3:00-3:15 PM

The Length of the Receiver Operating Characteristic (ROC) Curve and the Two Cutoff Youden Index Through a Robust Framework of Biomarker Discovery with Applications to High-Throughput Data

Leonidas Bantis, University of Kansas Medical Center

During biomarker discovery, high throughput technologies allow for simultaneous input of thousands of biomarkers that attempt to discriminate between healthy and diseased subjects. In such cases, proper ranking of biomarkers is highly important. Common measures, such as the area under the receiver operating characteristic (ROC) curve (AUC), as well as affordable sensitivity and specificity levels, are often taken into consideration. Strictly speaking, such measures are appropriate under a stochastic ordering assumption, which implies that higher measurements are more indicative of the disease. Such an assumption may lead to the rejection of useful biomarkers at this early discovery stage. We explore the length of a smooth ROC curve as a measure for biomarker ranking, which is not subject to directionality. We show that the length corresponds to a divergence and is identical to the corresponding length of the optimal (LR) ROC curve. We explore the relationship between the length measure and the AUC of the optimal ROC curve. We then provide a complete framework for the evaluation of a biomarker in terms of sensitivity and specificity and cutoffs.

3:15-3:30 PM

Quantile Function Regression Analysis for Adolescent Physical Activity Distributions in the Presence of Missing Data and Zero Inflation

Benny Ren, University of Pennsylvania

Wearable devices are increasingly used to obtain measurements in biomedical studies, producing high dimensional data. As opposed to the typical approach of modeling simple summaries that do not capture all information in these rich data, in this paper we introduce a statistical modeling framework to model subject-specific empirical distributions and assess how these vary across covariates. This paper adds to recent literature on distributional regression, including Wasserstein methods. Drawing on Wasserstein geometry, we propose a quantile function regression that accounts for missing data patterns, nonparametric covariate effects and zero inflation, which are important features in these studies. We incorporate a mixture model factorization to account for missing data patterns and a hurdle model factorization to account for zero inflation. We present a Bayesian computational and inferential framework that produces joint credible bands and inferential summaries that adjust for multiple testing. We use our framework to evaluate how activity distributions of a large cohort of adolescents vary with age, sex, and BMI to illustrate the benefits of this modeling strategy.

Tuesday, March 21, 2023 | 3:45-5:30 PM

73. RECENT ADVANCES IN PRECISION MEDICINE WITH HIGH-DIMENSIONAL BIOMEDICAL DATA

Organizer/Chair: Ziyi Li, The University of Texas MD Anderson Cancer Center

3:45-4:10 PM

Accounting for Network Noise in Graph-Guided Bayesian Modeling of High-Dimensional -Omics Data

Qi Long, University of Pennsylvania

High-dimensional omics data offer great promise in advancing precision medicine. Knowledge-guided statistical methods for analysis of omics data that can incorporate biological knowledge represented by graphs such as functional genomics have been shown to improve variable selection and predication accuracy and yield biologically more interpretable results, they typically use biological graph extracted from existing databases which is known to be incomplete and contain false edges. To address this issue, we propose a new knowledge-guided Bayesian modeling framework that treats the true biological graph as unknown or latent. Our model uses an adaptive structured shrinkage prior to incorporate the latent true biological graph to facilitate variable selection, and another set of priors motivated by the latent scale network model to connect two sources of noise-contaminated graph data, namely, biological graph extracted from a database and estimated covariance matrix for covariates, to the latent true graph. We develop an efficient MCMC algorithm for posterior sampling. We demonstrate the advantages of our model in simulations, and analysis of an AD genomics dataset.

4:10-4:35 PM

Integrated Reference-Informed Segmentation for Spatial Domain Detection in Spatial Transcriptomics

Xiang Zhou, University of Michigan

Detecting spatial domains on the tissue is an important step for characterizing the tissue transcriptomic landscape in spatial transcriptomics studies. While several statistical methods have been developed for detecting spatial domains, almost all of them have been focused on analyzing only one tissue section at a time from spatial transcriptomics. However, spatial transcriptomic studies often collect multiple adjacent sections from the same tissue or collect tissue samples from multiple individuals. Modeling multiple tissue sections together can thus borrow information across samples to potentially enhance the performance of spatial domain detection. In addition, and perhaps more importantly, existing single cell RNAseq data from the parallel research field can also provide important cell type specific expression information that could be used to complement spatial transcriptomics for accurate domain detection. Here, we develop a new method to leverage existing scRNA-seq datasets to facilitate domain detection on multiple tissue sections in spatial transcriptomics. We demonstrate the effectiveness of the method in both simulations and multiple real data applications.

4:35-5:00 PM

Leveraging Spatial Transcriptomics Data to Recover Cell Locations in Single-Cell RNA-Seq

Mingyao Li, University of Pennsylvania

Investigating the spatial origin of cells in solid tissues is essential for understanding the spatial organization of tissues and cell-cell communications. While scRNA-seq has made it possible to characterize cell types and states at an unprecedented resolution, the lack of physical relationships among cells has limited its applications when cell location information is needed. Spatial transcriptomics provides complementary information to single-cell data of dissociated cells/nuclei regarding the relationships between gene expression and spatial locations. Since a large amount of scRNA-seq data have been generated, it is desirable to recover their location information by utilizing the gene expression-spatial location relationship learned from spatial transcriptomics. In this talk, I will present methods that we recently developed to recover the spatial location information for cells in scRNA-seq at multiple levels, including 2D location as well as the spatial domain or tissue layer of a cell. We also provide uncertainty estimates for the recovered location information. I will show applications of these methods to data generated from mouse and human brains.

5:00-5:25 PM

Single-Cell and Spatial Omics in Dissecting Tumor Microenvironment

Linghua Wang, The University of Texas MD Anderson Cancer Center

The tumor microenvironment (TME) is a highly heterogeneous milieu consisting of phenotypically and functionally diverse immune and stromal cell populations and the cellular compositions and functional phenotypes of the TME can change from time to time during disease progression and the development of resistance to therapy. Single-cell and spatially resolved transcriptomics and immune repertoire sequencing are powerful technologies, providing unprecedented resolution to examine the highly complex tumor ecosystems, decipher the cellular and molecular landscapes of the TME, as well as characterize the dynamic interactions between cancer and TME cells. In this talk, Dr. Wang will present their most recent single-cell and spatial studies, she will share their novel discoveries and experience in profiling tumor immune microenvironment to advance our understanding of early tumorigenesis, immunotherapy response and metastasis, as well as toxicity associated with immune checkpoint therapies. She will conclude the talk with ongoing computational challenges & perspectives in the field.

74. RECENT ADVANCES IN STATISTICAL METHODS FOR NEURODEGENERATIVE DISEASE RESEARCH

Organizer/Chair: Panpan Zhang, Vanderbilt University Medical Center

3:45-4:10 PM

N-of-1 designs in Alzheimer's disease

Rebecca Betensky, NYU School of Global Public Health

Alzheimer's disease (AD) clinical trials require huge sample sizes and many years of follow-up, they are built on narrow concepts of AD and they do not acknowledge heterogeneity among patients. Given that only two drugs have been approved for AD in the past 20 years, it is clear that more flexible designs are needed to accelerate progress. One option is a rigorous formulation of a single patient trial, i.e., an N-of-1 trial. This automatically adjusts for some heterogeneity and confounding since subjects serve as their own controls. Although N-of-1 trials have been conducted in AD and other clinical settings, large parallel group RCTs remain the gold standard for drug development. I introduce a novel statistical design for single- and multi-person N-of-1 trials that incorporates sequential monitoring to increase efficiency by allowing for early stopping as soon as an optimal treatment -- or futility -- is identified for an individual or the population.

4:10-4:35 PM

Bridging Statistical Strategies for Censored Covariates to Neurodegenerative Diseases

Tanya Garcia, University of North Carolina Chapel Hill

Diseases of aging are expected to affect 153 million individuals worldwide by 2050. Treatments to slow these diseases will significantly decrease the projected impact, and modeling how disease symptoms worsen over time---the symptom trajectory---before and after a diagnosis can help evaluate if a treatment can slow the disease. Yet modeling the symptom trajectory is not easy because these diseases of aging progress slowly over decades, so studies that track symptoms often end before a diagnosis can be made. This makes time to diagnosis right-censored, leaving researchers with the challenge of trying to model the symptom trajectory without full information about when diagnosis occurs. Tackling this problem by modeling time to diagnosis has long been thought to be the best strategy, but when those models are even slightly wrong, that strategy produces biased results and incorrectly powered clinical trials. This talk presents practical, model-free strategies for this problem. The results can help produce robust estimates of the disease symptom trajectory which can assist in designing clinical trials that are powered to detect whether an experimental treatment is working.

4:35-5:00 PM

Multi-Layer Exponential Family Factor Models for Integrative Analysis and Learning Disease Progression

Yuanjia Wang, Department of Biostatistics, Columbia University

Current diagnosis of neurological disorders often relies on late-stage clinical symptoms, which poses barriers for developing effective interventions at the premanifest stage. Recent research suggests that biomarkers and subtle changes in clinical markers may occur in a time-ordered fashion and can be used as indicators of early disease. In this paper, we tackle challenges to leverage multi-domain markers to learn early disease progression of neurological disorders. We propose to integrate heterogeneous types of measures from multiple domains (e.g., discrete clinical symptoms, ordinal cognitive markers, continuous neuroimaging and blood biomarkers) using a hierarchical Multi-layer Exponential Family Factor (MEFF) model, where the observations follow exponential family distributions with lower-dimensional latent factors decomposed into shared factors across multiple domains and domain-specific factors. The MEFF model also captures nonlinear trajectory of disease progression and orders critical events of neurodegeneration measured by each marker. We apply the developed method to integrate biological, clinical and cognitive markers for Parkinson’s disease.

5:00-5:25 PM

Harnessing Latent Heterogeneity for Genetic Variants Underlying Alzheimer's Disease

Yongzhao Shao, New York University Grossman School of Medicine

Alzheimer’s disease is a major neurodegenerative disorder with considerable unmeasurable (latent) heterogeneity in etiology, clinical manifestation and prognosis. This study introduces a powerful genetic association test in the presence of latent (unmeasured) population heterogeneity. A novel latent class regression (LCR) approach is introduced to prioritize the large number of newly identified variants as targets for functional mechanistic studies and novel therapies. The flexible LCR analysis can identify desired subgroups among heterogeneous patients to facilitate the design and conduct of clinical trials to develop new targeted interventions and improve personalized patient care. The pipelines and algorithms are demonstrated using GWAS datasets from the Alzheimer’s Disease Neuroimaging Initiatives (ADNI). This talk is based on joint work with Dr. Linchen He and Dr. Yian Zhang at New York University Grossman School of Medicine.

75. WHEN MACHINE LEARNING MEETS MISSING DATA ANALYSIS

Organizer: Jiwei Zhao, University of Wisconsin-Madison
Chair: Fei Xue, Perdue University

3:30-4:10 PM

Semiparametric Estimation with Neural Networks for the Nonignorability Index

Jiwei Zhao, University of Wisconsin-Madison

In statistical data analysis, the assumptions imposed on nonignorable missing data are usually untestable. Sensitivity analysis is a statistical procedure to assess the discrepancy of the results between models with and without assuming the missingness is nonignorable. But this approach can involve complicated modeling and arduous computation, and can yield results that are highly sensitive to untestable model assumptions. Alternatively, nonignorability index, to measure the potential impact of nonignorability on an analysis, can be interpreted in terms of an intuitive parameter that captures the extent of sensitivity. In this paper, we propose semiparametric estimation for the nonignorability index, where the nonparametric components are estimated using neural networks, a flexible machine learning approach. The proposed estimator achieves the semiparametric efficiency bound in theory. Numerically, we conduct comprehensive simulation studies to evaluate its finite-sample performance and also apply it in a real data example for demonstration.

4:10-4:35 PM

On the Indispensability of Causality for Robust and Reliable Machine Learning Algorithms

Karthika Mohan, Oregon State University

Almost all modern-day AI and ML algorithms rely on strong assumptions such as data being IID, missingness being random and absence of unobserved confounders. While these assumptions make analysis relatively uncomplicated, they rarely hold true in real world datasets. For example, whether or not you contract an infectious disease is not determined exclusively by your vaccination status, it also depends on the vaccination status of people you interact with (i.e., data are not IID). If a variable such as income has several missing values, then missingness may not be random. Ignoring the true data generating process and relying on convenient assumptions is risky since it can potentially bias the research outcome. In this talk I will (i) outline how causal graphs can be used to model the data generating process, (ii) present conditions under which consistent estimates of quantities of interest such as causal effects can be computed, (iii) enumerate testable implications of a causal model and (iv) exemplify how bias can be detected, quantified and removed in non-iid datasets.

4:35-5:00 PM

Causal and Counterfactual Views of Missing Data Models

Razieh Nabi, Rollins School of Public Health, Emory University

It is often said that the fundamental problem of causal inference (CI) is a missing data problem -- the comparison of responses to two hypothetical treatment assignments is made difficult because for every experimental unit only one potential response is observed. In this talk, we consider the implications of the converse view: that missing data problems are a form of CI. We make explicit how the missing data problem of recovering the complete data law from the observed data law can be viewed as identification of a joint distribution over counterfactual variables corresponding to values had we (possibly contrary to fact) been able to observe them. Drawing analogies with CI, we show how identification assumptions in missing data can be encoded in terms of graphical models defined over counterfactual and observed variables. We note interesting similarities and differences between missing data and CI theories. The validity of any identification or estimation result relies on the assumptions encoded by the graph holding true. Thus, we also provide new insights on the testable implications of a few common classes of missing data models, and design goodness-of-fit tests for them.

5:00-5:25 PM

Are Deep Learning Models Superior for Missing Data Imputation in Surveys? Evidence from an Empirical Comparison

Fan Li, Duke University

Multiple imputation by chained equations (MICE) is one of the most widely used missing data imputation algorithms, but it is computationally intensive. Recently, missing data imputation methods based on deep learning models have been developed with encouraging results in small studies. However, there has been limited research on evaluating their performance in realistic settings compared to MICE. We conduct extensive simulation studies based on a subsample of the American Community Survey to compare the repeated sampling properties of four machine learning based MI methods: MICE with classification trees, MICE with random forests, generative adversarial imputation networks, and multiple imputation using denoising autoencoders. We find the deep learning imputation methods are superior to MICE in terms of computational time. However, with the default choice of hyperparameters in the common software packages, MICE with classification trees consistently outperforms, often by a large margin, the deep learning imputation methods in terms of bias, mean squared error, and coverage under a range of realistic settings.

76. COLLABORATING IN TEAM SCIENCE: NEGOTIATING PERSONALITIES, MENTORING NEW RESEARCHERS, AND OH, HOW DO I GET PROMOTED?

Organizer: Cyra Christina Mehta, Emory University School of Medicine
Chair: Brian Millen, Eli Lilly and Company

Panelists:
Portia D. Exum, SAS Institute

Renee' H. Moore, Drexel University Dornsife School of Public Health
Sowmya R. Rao, Boston University School of Public Health
Sean L. Simpson, Wake Forest University School of Medicine

Collaborative statisticians are team scientists who regularly interface closely with domain experts in other fields. Often, they collaborate on numerous projects, thoughtfully contributing to study design and applying appropriate statistical methods for the domain research question. Despite strong statistical training, skills needed to navigate personalities, mentor domain collaborators who may be experienced clinicians but naïve researchers, and balance career considerations are often learned through hands-on experience. The diverse panelists are from different sectors of academia and industry; they will share their journey and provide advice on how to create and manage successful interdisciplinary collaborations while making choices that enhance their career development and promotion.

77. POST GWAS: EPIGENOMICS, TRANSCRIPTOMICS, METABOLOMICS, AND CELL TYPES

Organizer/Chair: Jiebiao Wang, University of Pittsburgh

3:45-4:10 PM

Integrating Single-Cell Transcriptomic and Epigenomic Data with GWAS Summary Statistics to Prioritize Trait-Relevant Cell Types

Yuchao Jiang, University of North Carolina at Chapel Hill

Over the past decade, single-cell technologies have enabled a new era of high-resolution interrogation of cell-type diversity, vastly expanding our understanding of the role that cell types play in development and disease. Meanwhile, genome-wide association studies (GWAS) have successfully yielded genetic variants associated with various complex traits. For many complex traits, however, the specific cell types leading to risk are unknown. How should GWAS summary statistics be integrated with single-cell data to prioritize trait-relevant cell types? We propose statistical and computational methods to relate large-scale GWAS summary statistics to cell-type-specific gene expression and chromatin accessibility measurements from single-cell RNA and ATAC sequencing data. We use known trait-relevant tissues and cell types as ground truths for benchmark, adopt independent GWAS and single-cell datasets for enhanced rigor and reproducibility, and refer to PubMed keyword search and existing case-control studies for validation. The integrative analysis helps elucidate the underlying cell-type-specific disease etiology and prioritize important risk variants.

4:10-4:35 PM

DeepPerVar: A Multimodal Deep Learning Framework for Functional Interpretation of Genetic Variants in Personal Genome

Li Chen, University of Florida

Understanding the functional consequence of noncoding genetic variants is challenging. Genome-wide association studies or quantitative trait locus analyses may be subject to limited statistical power and linkage disequilibrium, and thus are less optimal to pinpoint the causal variants. By leveraging paired whole genome sequencing data and epigenetic functional assays in a population study, we propose a multi-modal deep learning framework to predict genome-wide quantitative epigenetic signals by considering both personal genetic variations and traits. The proposed approach can further evaluate the functional consequence of noncoding variants on an individual level by quantifying the allelic difference of predicted epigenetic signals. By applying the approach to the ROSMAP cohort studying Alzheimer’s disease (AD), we demonstrate that the proposed approach can accurately predict quantitative genome-wide epigenetic signals and in key genomic regions of AD causal genes, learn canonical motifs reported to regulate gene expression of AD causal genes, improve the partitioning heritability analysis, and prioritize putative causal variants in a GWAS risk locus.

4:35-5:00 PM

Accurate, Powerful, and Scalable Methodology for Metabolomic Data with Non-Ignorable Missing Observations and Latent Factors

Chris McKennan, University of Pittsburgh

It is well known that metabolomics data are corrupted by non-ignorable missing metabolite observations and latent confounding factors. While tempting to use models that incorporate both observed and missing data to facilitate statistically optimal estimators, the egregious non-normality of latent factors precludes the use of existing missing data models. Even if models were correct, estimation is intractable and prohibits phenome- and genome-wide analyses. Therefore, to make analyses manageable and attempt to glean something from the data, nearly all analysis pipelines first impute missing observations and subsequently perform analyses assuming complete data. While practical given the tools currently available, such imputation begets unreliable inference, thereby discrediting scientific conclusions. To address this critical methodological gap, we developed MS-NIMBLE, a provably accurate, powerful, and computationally efficient suite of methods that avoid invalid imputation to perform phenome-wide differential abundance analyses, metabolite genome-wide association studies (mtGWAS), and factor analysis with non-ignorable missing data.

5:00-5:25 PM

SR-TWAS: Leveraging Multiple Reference Panels to Improve TWAS Power by Ensemble Machine Learning

Jingjing Yang, Emory University

Although existing TWAS methods assume a single reference panel of transcriptomic and genetic data, multiple reference panels of a given tissue often exist. Thus, we developed the Stacked Regression based TWAS (SR-TWAS) tool to form optimal linear combinations of expression imputation models of the same tissue type. SR-TWAS can leverage multiple reference panels with increased effective training sample sizes as well as multiple regression methods to account for unkown underlying genomic architecture. We applied SR-TWAS to reference panels of GTEx V8 and ROS/MAP, to conduct real TWASs of Alzheimer's disease (AD) and Parkinson's disease (PD) using the most recent GWAS summary datasets. SR-TWAS identified respective 11 independent significant TWAS risk genes for AD and PD, including 5 novels for AD and 6 novels for PD. SR-TWAS tool provides a useful resource for leveraging multiple transcriptomic databases and models to increase expression prediction accuracy and TWAS power.

78. IMPROVING ESTIMATION OF DIFFICULT TO MEASURE HEALTH OUTCOMES

Organizer: Staci Hepler, Wake Forest University
Chair: Lucy D'Agostino McGowan

3:45-4:10 PM

An Integrated Abundance Model for Estimating County-Level Prevalence of Opioid Misuse in Ohio

Staci Hepler, Wake Forest University

Opioid misuse is a national epidemic and a significant drug related threat to the United States. While the scale of the problem is undeniable, estimates of the local prevalence of opioid misuse are lacking, despite their importance to policy-making and resource allocation. This is due, in part, to the challenge of directly measuring opioid misuse at a local level. In this paper, we develop a Bayesian hierarchical spatio-temporal abundance model that integrates indirect county-level data on opioid-related outcomes with state-level survey estimates on prevalence of opioid misuse to estimate the latent county-level prevalence and counts of people who misuse opioids. A simulation study shows that our integrated model accurately recovers the latent counts and prevalence. We apply our model to county-level surveillance data on opioid overdose deaths and treatment admissions from the state of Ohio.

4:10-4:35 PM

Estimating Seroprevalence of SARS-CoV-2 in Ohio: A Bayesian Multilevel Poststratification Approach with Multiple Diagnostic Tests

David Kline, Wake Forest University School of Medicine

In July 2020, there was great uncertainty around the spread of SARS-CoV-2. Despite its vital importance for public health policy, knowledge about the cumulative incidence of past infections was limited by challenges with diagnostic testing and the presence of mild or asymptomatic cases. Within this environment, competing narratives emerged around the prevalence of past SARS-CoV-2 infections, which would have had differing policy implications. To address this in July 2020, a population-representative household survey collected serum for SARS-CoV-2 antibody detection in Ohio in the United States. This talk describes a Bayesian statistical method developed to estimate the population prevalence of past infections accounting for the low positive rate; multiple, poor-quality, potentially correlated diagnostic tests; and the potential for non-ignorable non-response. This talk will also discuss the importance and potential practical challenges of conducting population-based surveillance surveys.

4:35-5:00 PM

Latent Class Models for Prevalence Estimation of Emerging Diseases using Verbal Autopsy Data

Zehang Li, University of California, Santa Cruz

Verbal autopsy (VA) is a widely-used tool to obtain information on causes-of-death when a medically certified cause-of-death is not available. It involves a structured questionnaire administered to family members or caregivers of a recently deceased person, and is widely adopted in many low- and middle-income countries without complete civil registration and vital statistics systems. During public health emergencies when new diseases emerge, VAs are usually the only feasible tool to gather information on cause-of-death in many settings. Existing statistical methods for analyzing VAs, however, are not suitable for such situations. We propose a novel latent class model framework to estimate the prevalence of a new disease using limited VAs collected under an informative selection process. We will also discuss extensions of the framework to monitor the dynamics of mortality rates over time.

5:00-5:25 PM

Methodological Considerations for Prevalence Estimation Using Multiple Systems Estimation in Small Areas

Katherine Thompson, University of Kentucky

Recently, multiple systems estimation has been used to address challenges in prevalence estimation when outcomes are difficult to observe. For example, the prevalence of outcomes that are difficult to observe (e.g., homelessness, opioid use disorder, or substance use disorders) have been widely underestimated due to limitations of existing data sources and difficulty in observing and recording cases. Capture-recapture methodology has been extended to improve opioid use disorder prevalence estimation at the county level in both Massachusetts and Kentucky. However, the application of such methods relies largely on assumptions about the underlying analysis data sets. For instance, smaller communities provide more instances of nonoverlapping data sources, requiring extensions of traditional capture-recapture methods to provide valid estimates. This talk will focus on the statistical aspects of implementing appropriate methods to account for nonoverlapping lists of subjects in small areas, and the results of choosing to ignore nonoverlapping lists in these types of analyses.

79. FOUNDATIONAL LEADERSHIP SKILL DEVELOPMENT FOR EARLY CAREER PROFESSIONALS

Organizers/Chairs: Veronica Bubb, Advance Research Associates; Claude Petit, Astellas

Panelists:
Kentaro Takeda, Astellas
Liwen Wu, Takeda
Kelly Huang, Advance Research Associates
Yu Du, Eli Lilly

The workforce is constantly reshaping business practices, and it is increasingly challenging to envision how the career path of others who came before us (Gen X or Baby Boomer) might differ from the career path that would apply to us in today’s fast-paced environment. However, what remains constant across generations is the fundamental principles of career progression. Post-pandemic in today's business climate, almost all of us have the option of remote work or a hybrid of remote and on-site. With the flexibilities that exist today, it is very much possible to have a work-life balance, a fulfilling job, and be able to develop your career at your chosen pace. Your career will be part of your life while you find success through the fundamentals of solid career building at an early stage. During this session, you will hear a review of these fundamentals from an industry executive and member of the Leadership in Practice Committee, and then a set of panelists of early career professionals will share their personal career journeys. The key takeaways from this session can be applied immediately and will enable you to jump-start your career.

80. CONTRIBUTED PAPERS: TIME-SERIES AND SPATIO/TEMPORAL MODELING

Chair: Matheus Bartolo Guerrero, King Abdullah University of Science

3:45-4:00 PM

Short-Term Forecast of a Pandemic

Cheng Cheng, St. Jude Children's Research Hospital

During a pandemic or epidemic, to effectively inform and support public health decision making in a local area (such as a county), it is important to have accurate short-term forecast of the disease case counts and/or resource usages (such as the number of hospitalizations). The sharp increases and decreases of daily case count in the COVID-19 pandemic has posed an analytical challenge to forecasting accurately the disease levels or resource usage in a locality; conventional forecast methods such as time series models cannot capture the sharp peaks. We have developed a nonparametric method which combines one-sided weighted average and a novel bias correction procedure. The weighted average is implemented by a one-sided discrete kernel with a tuning parameter that controls the span of the weights. The tuning parameter is determined by minimizing the average prediction errors using observed data in the recent past. Performance of this novel method was assessed by publicly available county-level daily COVID-19 case counts and a simulation study. The proposed method is able to adequately capture the sharp peaks and provide unbiased short-term forecast.

4:00-4:15 PM

A Group Testing Regression Model for Areal Data

Rongjie Huang, University of South Carolina

Group testing is widely used to reduce the costs of infectious disease surveillance. Rather than testing each specimen individually, multiple specimens are physically combined into a single pooled specimen, which is then tested the presence of an infectious agent or antibody. The testing cost reduction is achieved at the expense of data complication due to pooling effects and imperfect tests. In this paper, we present a Bayesian mixed effects regression model for group testing protocols for areal data with a conditional autoregressive (CAR) prior for the spatial random effects. Our model is suitable for all group testing protocols, including protocols with multiple testing assays. Provided the protocol involves retesting positive pools, the sensitivities and specificities of the assay(s) can be treated as unknown and estimated along with the other regression parameters. To illustrate the utility of our approach, we performed simulation studies for various scenarios and then applied our method to vector-borne disease surveillance data.

4:15-4:30 PM

A Subsampling Method Incorporated into the Bayesian Kriging Model

Sudipto Saha, Florida State University

Bayesian Kriging is a standard model in spatial statistics. However, one disadvantage of Bayesian Kriging is the computation time, which rapidly increases with the number of datapoints. The goal of this article is to allow a way for spatial statisticians to apply Bayesian Kriging to big datasets in a computationally feasible way, by incorporating a subsampling method into the Bayesian Hierarchical Model. Specifically, we extend the use of the recently introduced "data subset model" approach to spatial data, which has the advantage that one does not require any additional restrictive model assumptions. This provides a solution to computational bottlenecks that occur when applying Bayesian Kriging to ?big data?. We provide several properties of this new "spatial data subset model" approach in terms of moments, sill, nugget, and range under several sampling designs. We present the results of incorporating subsampling into Bayesian Kriging on several simulated datasets, and on a high-dimensional dataset consisting of 150,000 observations of daytime land surface temperatures as measured by the Terra instrument onboard the MODIS satellite.

4:30-4:45 PM

A Multivariate Spatio-Temporal Model for the Number of Imported COVID-19 Cases and COVID-19 Deaths in Cuba

Dries De Witte, KU Leuven, Belgium

To monitor the COVID-19 epidemic in Cuba, data on several epidemiological indicators have been collected. These data are available at the level of the municipalities, and can therefore be analyzed using spatio-temporal models. Univariate spatio-temporal models have been thoroughly studied, but when interest lies in studying the association between multiple outcomes, a joint model that allows for correlation between the spatial and temporal patterns is necessary. In this project, we propose a multivariate spatio-temporal model to study the association between the weekly number of COVID-19 deaths and the weekly number of imported COVID-19 cases in Cuba during 2021. To take into account the correlation between the spatial patterns, a multivariate conditional autoregressive prior (MCAR) is used. Correlation between the temporal patterns is taken into account by using two approaches; a multivariate random walk prior (1) or a multivariate conditional autoregressive prior (MCAR) (2). Separate spatially unstructured random effects and spatio-temporal interactions are also included in all models. All models were fitted using a Bayesian approach.

4:45-5:00 PM

Gaussian Processes for Irregular and Time-Lagged Time Series

Didong Li, University of North Carolina at Chapel Hill

Exploring the relationship, especially similarity, among multiple time series is a common question in many areas. These data are often irregular, and have time lags across features. To our best knowledge, no models have been specifically designed for this type of data. Common techniques i) fail to address the above two problems simultaneously, ii) may not have good performance for data with a small number of time points, and iii) do not leverage the correlation across time series when imputing into regular time points, leading to sub-optimal results. To address these challenges, we propose a model based on the Gaussian process, which leverages correlation among multiple time series, instead of fitting features separately, to achieve higher accuracy. Our method is flexible to deal with time-lagged, and irregular time series. Furthermore, our method produces interpretable parameters and can handle various usages, including ranking or clustering time series, interpolation, and forecasting with uncertainty quantification. Moreover, we prove the proposed kernels are valid and kernel parameters are identifiable. Our model yields improvements in a number of simulated settings and real applications for enhancer-promoter connection prediction compared to other state-of-the-art methods.

5:00-5:15 PM

A Bayesian Multivariate Mixture Model for Clustering Spot by Integrating H&E Image with Spatial Transcriptomics Data

Ye Seul Jeon, Yonsei University

High throughput spatial transcriptomics (HST) is a rapidly growing class of experimental methods that allows for gene expression profiling in tissue samples while maintaining the geographical position of each sequencing unit inside the tissue sample. We aim to identify cell subpopulations within a tissue sample by integrating H&E images with HST. First, we convert the image dataset into spatial data points to obtain density information about the spots. Based on this density infroation, we estimate the spot clustering membership and apply the spot membership to the Bayesian mixture model as prior information. Using publicly accessible human brain HST data, our approach recovers finely labeled brain layers more effectively than previous techniques.

5:15-5:30 PM

Bayesian Group-Shrinkage Based Estimation in Panel Var Models with Mixed Frequency Data

Nilanjana Chakraborty, University of Pennsylvania

Panel vector auto-regressive models are effective tools for modeling the evolution of multivariate time series across different regions. A key objective in this setting is to link the region-specific VAR models through appropriate homogeneity restrictions on their transition matrices to borrow strength from the common features, but also providing enough flexibility for regional idiosyncrasies. For macroeconomic data, this challenge is further enhanced by the fact that some variables are observed at a different frequency than others, and panel VAR literature for this mixed frequency setting is very sparse. We develop a Bayesian approach for mixed frequency panel VAR models that uses group shrinkage to borrow strength across regions and to tackle parameter proliferation. Existing Bayesian approaches for linking region-specific VARs try to fuse relevant coefficients of regional VAR transition matrices to a common value, while we employ a less stringent generalized hierarchical group-lasso prior based on groups of coefficients. Extensive performance evaluation on simulated data and on a motivating Eurozone macroeconomic dataset illustrates the efficacy of the proposed method.

81. CONTRIBUTED PAPERS: HIGH DIMENSIONAL, MULTIVARIATE, AND MISSING DATA METHODS

Chair: Xiaoqian Liu, University of Texas MD Anderson Cancer Center

3:45-4:00 PM

Multivariate Conditional Independence Testing for Microbial Network Construction

Hongjiao Liu, University of Washington

Microbial association networks help elucidate the complex interrelationships among different taxa within microbial communities. A key step in microbial network construction is to assess the pairwise conditional dependence between microbial taxa. While common approaches for evaluating conditional dependence treat individual taxa as univariate variables, a multivariate approach could be advantageous when the microbial taxa are each composed of multiple sub-taxa, e.g., a genus composed of multiple species. Here we propose a multivariate testing framework, Conditional RV, to assess the conditional dependence between two multivariate microbial features. In simulation studies, we show that Conditional RV has a better performance in recovering the true network compared to univariate competing methods, especially when heterogeneous relationships are present among the sub-taxa. When applied to a real vaginal microbiome data set, Conditional RV is able to produce a network consistent with existing knowledge on vaginal microbiota.

4:00-4:15 PM

High-Dimensional Rank-Based Inference for Individualized Treatment Effect in Single Index Varying Coefficient Model

Yishan Cui, Indiana University, Purdue University

Individualized treatment rule (ITRs) provides important information for patients' decisions on treatment based on their individual specific covariates. However, inference for ITRs could be difficult in high-dimensional situations when the interaction between treatment and other covariates is non-parametric. Estimation of the non-parametric function might induce non-negligible bias to the estimator of the ITRs. In this paper, we propose a rank-based inference procedure for ITRs under the semi-parametric single-index varying coefficient model when the non-parametric coefficient function is assumed to be monotone increasing. To avoid the estimation of the non-parametric function, the proposed method is based on maximum rank correlation. For testing purposes, the asymptotic distribution of the proposed estimator is derived using the de-biasing techniques. The finite sample performance of the proposed test is evaluated via Monte Carlo experiments as well as real data applications.

4:15-4:30 PM

Explainable Deep Learning Links Cancer Microbiome to Immune, Clinical, and Molecular Features of Cancer Patients

Sen Yang, Southern Methodist University

The relationship between the microbiome and cancer is becoming increasingly well established, yet the precise role of the microbiome in cancer remains unclear and requires further research. By investigating how the microbiome influences various immune, clinical, and molecular features, we can gain a better understanding of its impact and develop new diagnostic and therapeutic strategies. This highlights the need for computational methods and machine learning algorithms to analyze large amounts of microbiome data and identify potential microbial biomarkers. In response to this need, we developed the three-stage framework MB-LRP, which utilizes explainable deep learning models to discover bacterial biomarkers for immune, clinical, and molecular features of cancer patients. The framework includes building predictive models, identifying microbial biomarkers using layer-wise relevance propagation (LRP), and conducting enrichment analysis to validate LRP-identified microbial biomarkers. MB-LRP provides a comprehensive approach for detecting microbial biomarkers and has the potential to make significant advancements in cancer research and precision medicine.

4:30-4:45 PM

Multiple Imputation in High-Dimensional Data with Steady State Gibbs Sampler

Mario Keko, Jiann-Ping Hsu College of Public Health

The presence of the missing data is often an issue in the applications related to Biostatistics, Epidemiology, and Social Sciences fields. Literature has shown that the analysis without properly addressing this problem may produce biased results and be an inefficient use of the data. Multiple Imputation has been proposed as a solution with the literature showing that it can properly account for the missingness and outperform complete case analysis, which makes it a popular strategy for handling missing data. When it comes to high-dimensional data, the implementation of Multiple Imputation procedure becomes increasingly more complex, both from a theoretical and computational perspective. Several regularization methods and Bayesian Lasso regression for Multiple Imputations have been proposed and investigated. These methods often outperform the standard imputation approaches, but they may become computationally intensive. In our work, we are exploring more efficient regularization methods and compare and make conclusions regarding their efficiency and computational cost.

4:45-5:00 PM

Pan-Cancer Drug Response Prediction Using Integrative Principal Component Regression

Qingzhi Liu, Department of Biostatistics, University of Michigan-Ann Arbor

Precision oncology implementation strategies are in a strong need of integrative machine learning models to fill the gap between these different but related model systems. We develop an Integrative Principal Component Regression model (iPCR) to extract both joint variation across model systems for exploiting latent response predictors, and system-specific variation through a matrix decomposition method. Our iPCR has three benefits in the application of precision oncology: first, it quantifies the level of similarity and difference in gene expression data between cell lines and patient tumors; then, it matches cell lines with tumor samples based on the joint variation; and finally, it identifies key driver genes and pathways for treatment-specific response of pan-cancer patients. We apply iPCR on 12,000 patient samples from The Cancer Genome Atlas and TARGET programs across 30 tumor types, 1,400 cell lines from the Cancer Cell Line Encyclopedia, and drug responses of 16 cancer treatments. We show that iPCR performs favorably compared to four alternative approaches for patient drug response prediction in both real application and certain simulation scenarios.

5:00-5:15 PM

Robust Integration of Secondary Data Information into Main Outcome Analysis in the Presence of Missing Data

Daxuan Deng, Pennsylvania State University

In many clinical and observational studies, secondary data is often collected along with main data for each subject, which could help improve inference in main analysis but is rarely incorporated. In addition, missing data is commonly encountered in practice, which could lead to biased estimates if handled inappropriately. In this paper, we adopt empirical likelihood-based information borrowing method and inverse probability weighting (IPW) technique to improve estimation efficiency for the main analysis. Specifically, we propose a plug-in IPW estimator and show its equivalence to the standard joint estimator under mild conditions. The efficiency improvement is robust to misspecified working model in secondary analysis. Further, to handle bias due to missing secondary data, we propose a uniform mapping strategy that maps incomplete secondary data to a unified space by imputations. Extensive simulation shows that our methods are consistent, more efficient, and robust under various scenarios with missing data and/or secondary models misspecified. Finally, we illustrate our proposal by using the uniform data set from the National Alzheimer's Coordinating Center (NACC).

82. CONTRIBUTED PAPERS: MACHINE LEARNING METHODS/APPLICATIONS

Chair: Siddhesh Kulkarni, Bristol Myers Squibb

3:45-4:00 PM

Gaussian Differential Privacy and Its Applications

Weijie Su, University of Pennsylvania

Differential privacy has seen remarkable success as a rigorous and practical formalization of data privacy in the past decade. This privacy definition and its divergence-based relaxations, however, have several acknowledged weaknesses, either in handling composition of private algorithms or in analyzing important primitives like privacy amplification by subsampling. Inspired by the hypothesis testing formulation of privacy, this paper proposes a new relaxation, which we term `-differential privacy' (-DP). This notion of privacy has a number of appealing properties and, in particular, avoids difficulties associated with divergence-based relaxations. We conclude the talk with several applications of this new privacy framework to data-sensitive tasks.

4:00-4:15 PM

Multi-Class Classification for Multidimensional Functional Data Through Deep Neural Networks

Guanqun Cao, Auburn University

The intrinsically infinite dimension feature of the functional observations over multidimensional domains render the standard classification methods effectively inapplicable. To address this problem, we introduce a novel multiclass functional deep neural networks (mfDNN) classifier as a data mining and classification tool. Specifically, we consider sparse deep ReLU network reconstructions minimizing cross-entropy loss in the multi-class classification setup. Its neural network architecture allows us to employ modern computational tools in the implementation. The convergence rates of the misclassification risk functions are also derived for both fully observed and discretely observed multidimensional functional data. We demonstrate the performance of mfDNN on several benchmark datasets from various application domains.

4:15-4:30 PM

Multimodal Functional Deep Learning for Multi-Omics Data

Yuan Zhou, University of Florida

With rapidly evolving high-throughput technologies and ever-decreasing costs, it becomes feasible to collect diverse types of omics data in large-scale studies. While the multi-omics data generated from these studies hold great promise for innovative insights on biology mechanisms of human disease, the high-dimensionality of omics data and the complexity between various levels of omics data and disease phenotypes bring tremendous analytic challenges. To address these challenges and to facilitate ongoing multi-omics analysis, we propose a Multimodal Functional Deep Learning (MFDL) method for high-dimensional multi-omics data analysis. MFDL model the complex relationships between genetic variants and disease phenotypes through the hierarchical structure of deep neural networks and handle high-dimensional omics data by using the functional data analysis technique. Moreover, MFDL utilizes the structure of the multimodal model to model interactions between multi-omics data. Through simulation studies and real data applications, we demonstrate the advantages of MFNN in terms of prediction accuracy as well as being robust to the high dimensionality and noise of the data.

4:30-4:45 PM

An Association Test for Sequencing Data Based on Kernel Neural Networks

Tingting Hou, University of Florida

The recent development of artificial intelligence in genomics holds great promise to unravel the complex relationships between genetic variants and disease phenotypes and improve our understanding of the genetic etiology of complex diseases. However, due to the complexity of neural networks and their unknown limiting distributions, building significance tests on neural networks to examine complex genotype-phenotype relationships remains a great challenge. We previously developed a kernel-based neural network (KNN) method, which inherits features from both linear mixed models (LMM) and classical neural networks. Based on the KNN framework, we propose a Wald-type test to evaluate the joint association of a set of genetic variants with a disease phenotype, considering non-linear and non-additive effects. In addition, we also provide two tests to evaluate the linear genetic effect and non-linear/non-additive genetic effects (e.g., interaction effects), respectively. Through simulations, we demonstrated that our proposed method attained higher power compared to the sequence kernel association test (SKAT), especially in the presence of non-linear and interaction effects.

4:45-5:00 PM

Distillation Decision Tree

Xuetao Lu, The University of Texas MD Anderson Cancer Center

Black-box machine learning models are criticized as lacking interpretability, although they tend to have good prediction accuracy. Knowledge Distillation (KD) is an emerging tool to interpret the black-box model by distilling its knowledge into a transparent model. With well-known advantages in interpretation, decision tree is a competitive candidate of the transparent model. We name the decision tree generated from KD process as distillation decision tree (DDT). We discuss the theoretical foundations for DDT's structure stability which determines the validity of its interpretation. We prove that the structure of DDT can achieve stability under some mild assumptions. Meanwhile, we develop algorithms for stabilizing the induction of DDT, propose parallel strategies for improving the algorithm's computational efficiency, and introduce a marginal principal component analysis method for overcoming the curse of dimensionality in sampling. Simulated and real data studies justify our theoretical results, validate the efficacy of algorithms, and demonstrate that DDT can strike a good balance between the model's prediction accuracy and interpretability.

5:00-5:15 PM

Predicting Outcomes and Treatment Frequency Following Monthly Intravitreal Aflibercept for Macular Edema Secondary to Central Retinal Vein Occlusion: A Machine Learning Model Approach

Weiming Du, Regeneron Pharmaceuticals

Purpose: To develop machine learning (ML) models to predict visual and anatomic outcomes and treatment frequency in MEfCRVO patients at week 52 after intravitreal aflibercept injection (IAI). Methods: A dataset of 198 MEfCRVO patients treated with IAI 2 mg in the COPERNICUS/GALILEO trials was used. Patients were switched after 6 monthly IAI at week 24 to PRN dosing through Week 52. Random Forest was used to predict the absolute and change in best-corrected visual acuity (BCVA), patients with ?15-letter gain, absolute and change from baseline in central subfield thickness (CST) at Week 52, and IAI frequency during Weeks 24-52. Model performance was assessed using correlation coefficient (r) for continuous, and area under the curve (AUC) for categorical outcomes. Results: The ML model predicted the actual observed values with strong correlation for BCVA (r=0.87), change in BCVA (r=0.76), gain of ?15 letters (AUC=0.81), PRN injection frequency (AUC=0.83), and change in CST (r=0.76). There was no correlation between predicted and observed CST. Conclusions: ML successfully predicted visual and anatomic outcomes and dosing frequency with high accuracy, except for CST.

5:15-5:30 PM

An Optimal Transport and Graph Convolutional Based Transfer Learning Framework for Discovering High-Resolution Diagnostic and Prognostic Cell Types

Ziyu Liu, Purdue University

Traditional single-cell RNA sequencing approaches can correlate disease attributions with cell types but are constrained by the clustering resolution. To identify a more precise subpopulation of disease-related cell groups, we are developing novel deep transfer learning frameworks to relate single cell-level and patient-level data to derive disease association scores for every individual cell. Our first approach, DEGAS, mitigates the discrepancy between datasets by minimizing the Maximum Mean Discrepancy (MMD) distance between latent features from two data modalities. However, the MMD loss could face the vanishing gradients issue. Instead, the Optimal Transport (OT) theory provides a promising solution to the above problems. In our empirical simulation study, the OT-based model enables us to highlight disease-related cell sub-populations when the MMD loss fails to highlight these subgroups. Finally, we strengthen our framework by implementing a graph convolutional structure which is specifically designed for dealing with data with graph structure like spatial transcriptomics. We advocate for the use of deep transfer learning to identify high-risk cells in disease tissue.

83. CONTRIBUTED PAPERS: GENOME WIDE ASSOCIATION STUDIES

Chair: Teng Fei, Memorial Sloan Kettering Cancer Center

3:45-4:00 PM

BulkLMM: Fast Multiple Trait Genome Scans Using a Linear Mixed Model Approach with Permutation Tests

Zifan Yu, University of Tennessee Health Science Center

Linear mixed effects models (LMMs) are widely used in genome-wide association studies to account for the non-independence among subjects due to genetic relatedness. We introduce BulkLMM, a Julia package to perform fast genome scans for multiple traits and permutation testing using LMMs. Our goal is to perform near real-time expression quantitative trait locus (eQTL) scans in populations of modest size that would be suitable for web services such as GeneNetwork.org. Our approach uses easily parallelizable operations on modern multi-core computers. Running BulkLMM on mouse BXD spleen gene expression data with 79 individuals and 7321 genomic markers, we achieved a runtime of under a second for a univariate scan (single trait) with 1000 permutations. To scan all 35K gene expression traits, BulkLMM takes less than a minute. Our software speeds are comparable to that of GEMMA for a single trait without permutations. BulkLMM has the added benefit of permutation testing, efficiently scanning multiple traits, and integration with Julia for downstream analysis. (See BulkLMM package details on GitHub: https://github.com/senresearch/BulkLMM.jl)

4:00-4:15 PM

Integrating Multi-Omics Summary Data Using a Mendelian Randomization Framework

Chong Jin, New Jersey Institute of Technology

Using Mendelian randomization, we can identify the possible causal relationship between an omics biomarker and disease outcome using genetic variants as instrumental variables. This allows us to prioritize genes whose omics readouts can be used as predictors of the disease outcome through analyzing GWAS and QTL summary data. However, best practices are elusive when jointly analyzing the effects of multiple -omics biomarkers annotated to the same gene of interest. To bridge this gap, we propose powerful combination tests that integrate multiple correlated p-values without knowing the dependence structure between the exposures. Our simulation experiments demonstrate the superiority of our proposed approach compared with existing methods adapted to the setting of our interest. The method is illustrated on a multi-omics Alzheimer's disease dataset.

4:00-4:15 PM

Transfer Learning with Summary Statistics and Its Applications to Polygenetic Risk Score Prediction

Haotian Zheng, University of Pennsylvania

This paper considers the estimation and prediction of a high-dimensional linear regression in the settings of transfer learning where we only observe summary statistics in the auxiliary studies, together with external data for the estimation of linkage disequilibrium. We develop a method for estimation of the regression coefficient vector based on auxiliary summary statistics and external data. We show the improvement of estimation when we additionally have summary statistics of auxiliary studies, but the rate for estimation is slower than individual data transfer learning. The findings are illustrated through simulation studies using real genotype data, and an analysis of a large scale GWAS focusing on the polygenic risk score (PRS) predictions of blood related phenotypes of samples in the Penn Medicine Biobank, using summary statistics from UK Biobank as auxiliary studies.

4:30-4:45 PM

Testing for Interaction in Genome-Wide Association Studies: Surprising Statistical Properties

Huanlin Zhou, The University of Chicago

In a genome-wide association study (GWAS) for a trait, we consider the problem of testing for interaction, either between pairs of SNPs throughout the genome (epistasis) or between each SNP and a given environmental variable. We show that even in simplified models in which all SNPs are assumed independent and all individuals are assumed independent, in a GWAS to detect gene x gene or gene x environment interaction, the distribution of the genome-wide p-values under the null hypothesis of no interaction is not i.i.d. uniform as it would be for an ordinary genotype-trait association analysis in this simplified setting. This is a surprising result that is specific to detection of interaction in a GWAS, and it could explain why such detection has so far proved challenging and difficult to replicate. We have developed methods to correct the interaction test statistics to make their p-values have the expected null distribution. Our methods can be extended to the linear mixed model case to allow for covariates and population structure. Through simulation and data analysis, we demonstrate the severity of the problem and value of the new methods in resolving it.

4:45-5:00 PM

Inference of Causal Networks Using Bi-Directional Mendelian Randomization and Network Deconvolution with GWAS Summary Data

Zhaotong Lin, University of Minnesota

Inferring causal relationships among potential risk factors and diseases from observational data is both important and challenging, e.g. due to hidden confounding. A recently developed instrumental variable (IV) method called Mendelian randomization (MR) has been increasingly applied for causal inference with observational. However, the current practice of MR has been largely restricted to investigating the total causal effect between two traits, while it would be more useful to infer the direct causal effect between any two of many traits (by accounting for indirect or mediating effects through other traits). In this work, we first extend bi-directional MR-cML, robust to the violation of all three IV assumptions, to overlapping-sample scenario, then apply it to infer a causal network of total effects among multiple traits. Finally we apply graph deconvolution to infer a causal network of direct effects. We also develop strong theoretical support for our proposed method. Lastly, the application to 17 large-scale GWAS summary datasets including 11 risk factors and 6 diseases showcases the promising advantage of our method in studying causal pathways among multiple traits.

84. CONTRIBUTED PAPERS: STATISTICAL METHODS FOR IMAGE-BASED DATA ANALYSIS

Chair: Andrew Chen, University of Pennsylvania

3:45-4:00 PM

Improving Robustness of Radiomics Pipelines to Differences in Acquisition and Segmentation in Clinical Trial Data

Hannah Horng, Department of Bioengineering, University of Pennsylvania

Radiomics is a promising approach for guiding decision-making in clinical trials by leveraging existing databases of patient imaging. However, there are sources of variability in radiomics pipelines that could reduce reproducibility. These include differences in segmentation due to inter-rater variability and batch effects in the features caused by differences in image acquisition. While methods such as ComBat may address these sources of variability, they have not been tested on large clinical trial datasets. In this work, we develop and apply methods for improving robustness to technical variability in radiomics pipelines using computed tomography (CT) images collected as part of Merck?s large-scale pembrolizumab trial program. We apply ComBat to radiomic features extracted from the KEYNOTE-052/086 (n=472) and 001 (n=944) trials and show that it can improve robustness to differences in CT manufacturer using statistical testing for distributional differences due to batch and predictive modeling. We also develop a pipeline for simulation of differences in segmentation and identify features robust to these differences for feature selection.

4:00-4:15 PM

Bayesian Pathway Analysis over Brain Network Mediators for Survival Data

Xinyuan Tian*, Yale University

Technological advancements in noninvasive imaging techniques facilitate the construction of whole brain interconnected networks, known as brain connectivity. Existing approaches to analyze brain connectivity frequently disaggregate the entire network into a vector of unique edges, leading to a substantial loss of information. In this article, motivated by the need to explore the effect mechanism among the genetic exposure, brain connectivity and time to disease onset, we propose an integrative Bayesian framework to model the effect pathway between each of these components. To accommodate the biological architectures of brain connectivity, we develop a structural modeling framework including a symmetric matrix-variate accelerated failure time model, and a symmetric matrix response regression to characterize the effect paths. Extensive simulations confirm the superiority of our method compared with existing alternatives. By applying the proposed method on the landmark Alzheimer's Disease Neuroimaging Initiative (ADNI) study, we obtain neurobiologically plausible insights that may inform future intervention strategies.

4:15-4:30 PM

Spectral Topological Data Analysis for EEG Brain Signals

Anass El Yaagoubi Bourakna, King Abdullah University of Science and Technology

Topological data analysis has become a powerful approach over the last twenty years, mainly because of its ability to capture the shape and the geometry inherent in the data. Specifically, the use of persistence homology for analyzing functional brain connectivity has witnessed considerable success in the literature. It solves the problem of connectivity matrix thresholding at arbitrary levels by considering a filtration of the weighted network across all possible threshold values. Such approaches for analyzing the topological structure of functional brain connectivity rely on simple connectivity measures such as Pearson correlation. To overcome this limitation, we propose a frequency-specific approach that leverages coherence to assess the brain's functional connectivity, leading to a novel topological summary, the spectral landscape, which is an extension of the persistence landscape. Using this novel approach to analyze the EEG brain connectivity of ADHD subjects, we shed light on the frequency specific differences in the topology of brain connectivity between healthy controls and ADHD subjects.

Wednesday, March 22, 2023 | 8:30-10:15 AM

85. ANALYSIS OF DISTRIBUTED HEALTH DATA: NOVEL APPROACHES AND APPLICATIONS

Organizer: Lu Tang, University of Pittsburgh
Chair: Yiwang Zhou, St. Jude Children's Research Hospital

8:30-8:55 AM

Organizational Collaboration with Assisted Learning

Jie Ding, University of Minnesota

Humans develop knowledge from individual studies and joint discussions with peers, even though each individual observes and thinks differently. Likewise, in many emerging application domains, collaborations among organizations or intelligent agents of heterogeneous nature (e.g., health institutes, companies of different sectors, and government agencies) are often essential to resolving challenging problems that are otherwise impossible to be dealt with by a single organization. However, to avoid leaking useful and possibly proprietary information, an organization typically enforces stringent security measures, significantly limiting such collaboration. This talk will introduce a new research direction called "Assisted Learning" that aims to enable organizations to assist each other in a decentralized, personalized, and private manner without sharing data, models, or learning objectives. This includes new concepts and methods inspired by cross-disciplinary perspectives such as statistics, optimization, and information theory.

8:55-9:20 AM

Targeting Underrepresented Populations in Precision Medicine: A Federated Transfer Learning Approach

Rui Duan, Harvard University

The limited representation of minorities and disadvantaged populations in large-scale clinical and genomics research has become a barrier to translating precision medicine research into practice. Due to heterogeneity across populations, prediction models are often found to be underperformed in underrepresented populations, and therefore may further exacerbate known health disparities. We propose a two-way data integration strategy that integrates heterogeneous data from diverse populations and multiple healthcare institutions via a federated transfer learning approach, which improves the estimation and prediction accuracy in underrepresented populations, and reduce the gap in model performance across populations. Our theoretical analysis reveals how the estimation accuracy of the proposed method is influenced by communication budgets, privacy restrictions, and heterogeneity across populations. We demonstrate the feasibility and validity of our method through numerical experiments, and apply them to construct genetic risk prediction models using data from large-scale data networks.

9:20-9:45 AM

Avoiding Significance Fetishism in Meta-Analyses: Pitfalls and Reporting Recommendations

Maya Mathur, Stanford University

The published original literature is biased as a result of researchers' fetishizing statistical significance. Well-conducted meta-analyses and systematic replication studies have the potential to help mitigate this bias and to estimate its severity. However, as I will argue, standard approaches to analyzing and reporting on meta-analyses often themselves focus explicitly or implicitly on statistical significance. Using real examples, I will show how these practices have led to misleading conclusions, such as illusory conflicts between meta-analyses that in fact show little disagreement. I will discuss recently introduced statistical metrics and reporting approaches that shift the focus to effect sizes, and their potential heterogeneity, rather than on statistical significance alone.

9:45-10:10 AM

Fairness-Oriented Decision Learning in Distributed Research Networks

Lu Tang, University of Pittsburgh

This work introduces a robust individualized decision learning framework with for distributed research networks (DRNs) with the goal of ensuring fairness among sites in a DRN. The convention in decision learning for DRN is to maximize or minimize an aggregated loss function across sites. However, an aggregated loss does not provide fairness guarantees, hence it may disproportionately advantage or disadvantage some of the sites, resulting in poor local performance at certain sites. From a causal perspective, the proposed learning framework robustifies sites via a newly defined quantile- or infimum-optimal decision rule for improving the outcome of the worst-off site among all sites. The reliable performance of the proposed method is demonstrated through synthetic experiments and real-data applications.

86. CURRENT APPROACHES FOR ACCELERATING DRUG APPROVALS IN RARE DISEASE

Organizer/Chair: Brad Carlin, PharmaLex—US

8:30-8:55 AM

The FDA Accelerating Rare Disease Cures (ARC) Program: Opportunity for Innovation

Dionne Price, Food and Drug Administration

Rare disease settings present unique challenges for drug development and regulatory decision making. The challenges include, but are not limited to, small patient populations, phenotypic heterogeneity within diseases, lack of reliable and well-defined endpoints, and limited information on the natural history. These challenges represent significant opportunities for innovation in endpoint development, study designs, and statistical analyses. To drive innovation in rare diseases, the Center for Drug Evaluation and Research of the Food and Drug Administration launched the Accelerating Rare Disease Cures (ARC) Program in May 2022. This talk will provide an overview of the ARC program and some of the methodologic opportunities it affords.

8:55-9:20 AM

snSMART Design and Methods to Register a Drug in Rare Diseases

Kelley Kidwell, University of Michigan

A more recent rare disease trial design, the small n, sequential, multiple assignment, randomized trial (snSMART), is a promising tool to incorporate dose finding and confirmation of the dose effect in one trial. Motivated by disease settings such as isolated skin vasculitis and Duchene muscular dystrophy, this two-stage design is most appropriate for stable or slowly progressing rare diseases and may improve recruitment and retention over more commonly used rare disease trial designs. In this talk, we will introduce the snSMART design and corresponding Bayesian methods. We will show how external data can be formally incorporated into a robust exchangeable normal-normal hierarchical model to lessen the number of participants who receive placebo. Bias and efficiency of treatment effect estimators will be compared across various methods considering one stage of data, two stages of data and incorporation of external data.

9:20-9:45 AM

Bayesian Propensity Score Based Approaches for Leveraging Real World Data in Rare Disease Clinical Trials

Jian Zhu, Servier Pharmaceuticals

Using RWD and other external data to supplement trial data is particularly relevant in rare diseases. On one hand, Bayesian methods have been proposed to borrow such external data; on the other hand, with more accessible individual patient level data, propensity score methods such as stratification have been used to specifically balance baseline characteristics and prognostic factors across data sources. We explored and generalized a framework combining propensity score stratification and Bayesian borrowing methods to improve the estimation of the current trial's parameter of interest. Various Bayesian methods were explored, including double hierarchical prior, robust mixture prior and power prior. Common caveats and issues for similar framework will be explored along with proposed solutions. Finally, findings and conclusions based on an extensive simulation study will be discussed.

9:45-10:10 AM

DISCUSSANT

Shirin Golchi, McGill University

87. BUILDING BIOSTATISTICS CAPACITY IN SUB-SAHARAN AFRICA

Organizer: Michael Rosenblum, Johns Hopkins Bloomberg School of Public Health
Chair: Kelly Van Lancker, University of Ghent; Johns Hopkins University

8:30-8:55 AM

Biostatistics and Health Data Science Training in Sub-Saharan Africa: Challenges and Opportunities

Henry Mwambi, University of KwaZulu-Natal

Biostatistics training in SSA has been slow due to a number of reasons and or challenges. For a long period of time Statistics training in Africa has been skewed towards the theoretical approach with minimal application. Availability of resources such as software and training to undertake real world practical application, was another challenge that hampered the growth of biostatistics in the region. However in recent times a number of novel initiatives, have been implemented to help enhance biostatistics training in the region. This talk will highlight some of these developments and achievements realised so far. One of the notable initiative, that the talk will focus on significantly, is the formation of the Sub-Saharan Africa Consortium for Advanced Biostatistics training (SSACAB), which has become an important vehicle for enhancing biostatistics capacity in the region and by extension health data science. The talk will also highlight challenges that are still encountered, that need attention in order to accelerate biostatistics and health data science training and research in the region.

8:55-9:20 AM

Strengthening Biostatistics Resources in Sub-Saharan African Countries

Misrak Gezmu, National Institute of Allergy and Infectious Diseases, National Institutes of Health

In the last three decades, because of disease burdens (HIV, TB, and other communicable and non-communicable diseases) and increase in funding, biomedical research in sub-Saharan African (SSA) countries has increased. Most of the biomedical research projects are conducted with the research teams from North-South collaborative institutions. In most SSA countries, biomedical research is new and biostatistical resources are not well developed. Conducting biomedical research requires a research discipline that requires training, experience, and critical thinking. Strengthening local biostatistics resources will enable the local biomedical research teams to have statistical leaders that contribute to the critical thinking, and to the design and analysis of conducting the biomedical projects. I will discuss the summary of the 2009 workshop conducted in Bethesda, Maryland, and the subsequent follow up workshops conducted in Gaborone, Botswana (2011), and in Banjul, The Gambia (2022). Lessons learned from these workshops and the influence the workshops have in increasing the biostatistical activities in SSA countries will be assessed.

9:20-9:45 AM

Building and Sustaining a Kenyan/American Partnership in Biostatistics

Joseph Hogan, Brown University School of Public Health

Sub-Saharan Africa (SSA) has seen extensive growth in biomedical research and a rapid increase in data collection and infrastructure. Yet a considerable amount of biostatistical work is conducted in Europe and the US, a disparity that characterizes many 'north-south' collaborations in global health. Investments by NIH and others have partially alleviated this disparity, but substantial gaps remain. We will describe a long-standing research and training partnership between Moi and Brown Universities. Formed 15 years ago, it has been sustained by an NIH training grant, an evolving program of methods research, and a portfolio of collaborative research. The training program has produced 1 PhD and 6 MS degree recipients, nearly all of whom are working at Moi on funded research; conducted 5 training workshops for hundreds of professional statisticians in SSA; and motivated several statistical research projects. We will discuss navigating cultural and institutional differences, addressing challenges of 'global health colonialism', and essential principles that have been fundamental to the ongoing success of the partnership, and indicate plans for the future.

9:45-10:10 AM

Building Biostatistics Capacity in Sub-Saharan Africa

Michael Rosenblum, Johns Hopkins Bloomberg School of Public Health

I will present an ongoing project that aims to increase local capacity for conducting clinical trials in low- and middle-income countries (LMIC), with a focus on sub-Saharan Africa. The project will provide open-source software and training materials for clinical trial design and analysis for clinical investigators and statisticians in LMIC. This is part of a collaboration with investigators from 10 LMIC countries. Most of the collaborators are in the Sub-Saharan Africa Consortium for Advanced Biostatistics (SSACAB), which aims to improve biostatistical capacity in Africa according to the needs identified by African institutions). We will collaborate with investigators who are designing their first clinical trials by providing support in optimizing the trial design (specifically, improving precision using covariate adjustment), writing the statistical analysis plans, and grant writing. This project is one step toward the larger goal of creating a comprehensive set of training materials that covers all key aspects of the design and statistical analysis of clinical trials, focused on needs of trials led by investigators in LMIC. Our motivation is that we were told by SSACAB members that it would be useful to create more materials to train the next generation of clinical trials biostatisticians in Sub-Saharan Africa.

88. INDIVIDUALIZATION, MOBILIZATION, AND BEYOND

Organizer/Chair: Jiwei Zhao, University of Wisconsin-Madison

8:30-8:55 AM

Multi-Resolution Latent Factor Model for Wearable Device Data

Annie Qu, UC Irvine

Rich longitudinal data tracking people's physical activities and health status enable us to deliver non-invasive interventions in real time. We propose a latent dynamic model (LDM) for multi-resolution data integration in mobile health. The proposed method utilizes non-parametric latent dynamic factors to capture the underlying trend in longitudinal data with mixed resolution and irregular time intervals. Our numerical studies illustrate that the LDM is able to capture the dynamic trend of multi-resolution time-series data more accurately compare to existing methods. Importantly, the non-parametric latent factors is capable of recovering large portion of missing data occur frequently in mobile health studies. Our real data problem is motived by Garmin watch and Oura ring data monitoring caregivers's stress for caregiving dementia patients.

8:55-9:20 AM

Homogeneity Pursuit of Functional Association Parameter in Scalar-on-Function Regression Analysis of Physical Activity Data from Wearable Devices

Peter Song, University of Michigan

We consider a scalar-on-function regression analysis of physical activity data collected from a wearable device, in which the functional predictor is given by subject's Occupation-Time curve (OTC) that presents a proportional continuum of time spent at or above varying activity levels. We invoke a mixed integer optimization (MIO) paradigm to formulate a fused estimation method for homogeneity pursuit under the L0 penalization. This new approach can perform a simultaneous operation of changepoint detection and step-functional parameter estimation. We show through extensive simulation experiments that the proposed MIO methodology enjoys both estimation accuracy and computational efficiency. Under some mild regularity conditions, we establish a finite error bound for the changepoint selection consistency and parameter estimation consistency. We apply the proposed MIO method on a real-world data analysis to assess the influence of physical activity on biological aging.

9:20-9:45 AM

Fairness-Oriented Learning for Optimal Individualized Treatment Rules

Lan Wang, Miami Herbert Business School, University of Miami

A notable drawback of the standard approach to optimal individualized treatment rule (ITR) estimation is that the estimated optimal ITR may be suboptimal or even detrimental to certain disadvantaged subpopulations. Motivated by the importance of incorporating an appropriate fairness constraint in optimal decision making (e.g., assign treatment with protection to those with shorter survival time), we propose a new framework that aims to estimate an optimal ITR to maximize the average value with the guarantee that its tail performance exceeds a prespecified threshold. The optimal fairness-oriented ITR corresponds to a solution of a nonconvex optimization problem. Furthermore, we extend the proposed method to dynamic optimal ITRs. (Joint work by Ethan Fang, Zhaoran Wang and Lan Wang)

9:20-9:45 AM

Fairness-Oriented Learning for Optimal Individualized Treatment Rules

Lan Wang, Miami Herbert Business School, University of Miami

A notable drawback of the standard approach to optimal individualized treatment rule (ITR) estimation is that the estimated optimal ITR may be suboptimal or even detrimental to certain disadvantaged subpopulations. Motivated by the importance of incorporating an appropriate fairness constraint in optimal decision making (e.g., assign treatment with protection to those with shorter survival time), we propose a new framework that aims to estimate an optimal ITR to maximize the average value with the guarantee that its tail performance exceeds a pre-specialized threshold. The optimal fairness oriented ITR corresponds to a solution of a nonconvex optimization problem. Furthermore, we extend the proposed method to dynamic optimal ITRs. (Joint work by Ethan Fang, Zhaoran Wang and Lan Wang)

9:45-10:10 AM

Use of Mobile Data to Examine Compliance in Clinical Trials

Heping Zhang, Yale University

Compliance to interventions is very important for behavioral modification trials, but is generally difficult to monitor. We used data from a multi-center trial, called "Improving Reproductive Fitness through Pretreatment with Lifestyle Modification in Obese Women with Unexplained Infertility". We used the data from 358 participants 57,505 observations from this trial. We chose the daily total steps to define a measure of compliance. The trial was composed of multiple sequential phases, so we defined phase-specific compliance scores, and weighed the phase-specific scores to obtain the overall compliance score for each participant. We examined how the compliance scores were associated with other variables, and the association between pregnancy outcomes with the overall compliance score, while adjusting selected covariates. The overall compliance score was significantly associated with baseline variables including baseline steps, education level, history of prior conception, history of prior pregnancy loss, and baseline estradiol level, but not significantly associated with pregnancy outcomes.

89. ADVANCES IN CONTINUOUS-TIME CAUSAL INFERENCE WITH TIME-VARYING TREATMENTS

Organizer: Jinghao Sun, Yale School of Public Health
Chair: Forrest Crawford, Yale School of Public Health

8:30-8:55 AM

Causal Identification for Continuous-Time Stochastic Processes

Jinghao Sun, Yale University

Many real-world processes are trajectories that may be regarded as continuous-time "functional data". Examples include patients' biomarker concentrations and environmental pollutant levels. Corresponding advances in data collection have yielded near continuous-time measurements, from e.g. wearable digital devices and environmental sensors. Statistical methodology for estimating the causal effect of a time-varying treatment, measured discretely in time, is well developed. But discrete-time methods like the g-formula, structural nested models, and marginal structural models do not generalize easily to continuous time. Moreover, researchers have shown that the choice of discretization time scale can seriously affect the quality of causal inferences about the effects of an intervention. In this paper, we establish causal identification results for continuous-time treatment-outcome relationships for general cadlag processes under continuous-time confounding, through orthogonalization and weighting. We use concrete running examples to demonstrate the plausibility of our identification assumptions, and their connections to the discrete-time g methods literature.

8:55-9:20 AM

Causal Graphs for Identification of Causal Effects in Continuous-Time Event-History Analyses

Kjetil Røysland, Department of Biostatistics, University of Oslo

We consider continuous-time survival or more general event-history settings, where the aim is to infer the causal effect of a time-dependent treatment process. This is formalised as the effect on the outcome event of a (possibly hypothetical) intervention on the intensity of the treatment process, i.e. a stochastic intervention. To establish whether valid inference about the interventional situation can be drawn from typical observational, i.e. non-experimental, data we propose graphical rules indicating whether the observed information is sufficient to identify the desired causal effect by suitable re-weighting. In analogy to the well-known causal directed acyclic graphs, the corresponding dynamic graphs combine causal semantics with local independence models for multivariate counting processes. Importantly, we highlight that causal inference from censored data requires structural assumptions on the censoring process beyond the usual independent censoring assumption, which can be represented and verified graphically. Our results establish general non-parametric identifiability and do not rely on particular survival models.

9:20-9:45 AM

Continuous-time TMLE

Helene Charlotte Wiese Rytgaard, University of Copenhagen

Targeted learning (TMLE) is a general framework for semiparametric efficient substitution estimation of causal parameters that combines machine learning with asymptotic statistical inference. The continuous-time TMLE is a generalization of the targeted learning methodology for estimation of time-varying interventional effects in settings where interventions, covariates and outcome can happen at subject-specific points in time. In this talk, I will discuss different aspects of the continuous-time TMLE. The general approach utilizes a counting process framework, and involves construction of hazard-based initial estimators and an extension of the targeting procedure that combines updating steps of intensities and conditional expectations to solve the efficient influence curve equation. Applications include a large variety of problems in pharmacoepidemiology, where hazard ratios have historically been widely used to measure association between exposures and time-to-event outcomes but are by now known to not generally yield a causal interpretation.

9:45-10:10 AM

Estimating the Causal Effects of Multiple Intermittent Treatments with Application to COVID-19

Liangyuan Hu, Rutgers University

To draw real-world evidence about the comparative effectiveness of multiple time-varying treatment regimens on patient survival, we develop a joint marginal structural proportional hazards model and novel weighting schemes in continuous time to account for time-varying confounding and censoring. Our methods formulate complex longitudinal treatments with multiple ``start/stop'' switches as the recurrent events with discontinuous intervals of treatment eligibility. We derive the weights in continuous time to handle a complex longitudinal dataset on its own terms, without the need to discretize or artificially align the measurement times. We further propose using machine learning models designed for censored survival data with time-varying covariates and the kernel function estimator of the baseline intensity to efficiently estimate the continuous-time weights. An extensive simulation was conducted to investigate the operating characteristics of our proposed methods. We apply the proposed methods to a COVID-19 dataset to estimate the causal effects of several COVID-19 treatment strategies on in-hospital mortality or ICU admission.

90. NEW DEVELOPMENTS AND CHALLENGES FOR STATISTICAL GENETICS AND GENOMICS

Organizer/Chair: Zilin Li, Indiana University School of Medicine

8:30-8:55 AM

Ensemble Methods for Testing a Global Null in Whole Genome Sequencing Association Studies

Xihong Lin, Department of Biostatistics and Department of Statistics, Harvard University

Testing a global null is a canonical problem in statistics and has a wide range of applications. In view of the fact of no uniformly most powerful test, motivated by the success of ensemble learning methods for prediction or classification, we propose an ensemble framework for testing that mimics the spirit of random forests to deal with the challenges. Our ensemble testing framework aggregates a collection of weak base tests to form a final ensemble test that maintains strong and robust power. The key component of the framework is to introduce a certain random procedure in the construction of base tests. We then apply the framework to four problems about global testing in different classes of alternatives arising from Whole Genome Sequencing (WGS) association studies. The theoretical optimality is established in terms of Bahadur efficiency. Extensive simulations are conducted to demonstrate type I error control and power gain of the proposed ensemble tests. In an analysis of the WGS data from the Atherosclerosis Risk in Communities (ARIC) study, the ensemble tests demonstrate substantial and consistent power improvement compared to other existing tests.

8:55-9:20 AM

Knockoff-Based Statistics for the Identification of Putative Causal Loci in Genetic Studies

Iuliana Ionita-Laza, Columbia University

Knockoff-based methods are becoming increasingly popular due to their enhanced power for locus discovery and their ability to prioritize putative causal variants in a genome-wide analysis. However, because of the substantial computational cost for generating knockoffs, existing knockoff approaches cannot analyze biobank-scale datasets. I will discuss a scalable knockoff-based method for population-based designs, and related extensions to family-based designs. I will show applications to the UK biobank data, and some family-based studies for autism spectrum disorders.

9:20-9:45 AM

Integration of Histopathology Images with Genomics Data for Cancer Precision Medicine

Kun Huang, Indiana University School of Medicine

Histopathology images are essential for cancer diagnosis and prognosis while various omics data are used for subtyping and prediction. Integration of histopathology images with multi-omics data can lead to discoveries of new subtypes and biomarkers. To ensure interpretability, we developed a series of pipelines for extracting interpretable quantitative cellular and tissue morphological features from histopathology images. These features include nucleic shape (e.g., area, eccentricity, color), cell density (i.e., distances between neighboring cells as defined by Delaunay triangulation), and tissue proportion (e.g., ratio between segmented stromal and epithelial regions). Then we integrated these features with various omics and clinical data using a series of methods based on canonical correlation analysis and deep learning algorithms. We demonstrate our methods on multiple cancer types including breast and kidney cancers.

9:45-10:10 AM

Mendelian Randomization Mixed-Scale Treatment Effect Robust Identification and Estimation for Causal Inference

Zhonghua Liu, Department of Biostatistics, Columbia University

Standard Mendelian randomization (MR) analysis can produce biased results if the genetic variant defining an instrumental variable (IV) is confounded and/or has a horizontal pleiotropic effect on the outcome of interest. We provide novel identification conditions for the causal effect of a treatment in the presence of unmeasured confounding by leveraging a possibly invalid IV for which both the IV independence and exclusion restriction assumptions may be violated. The proposed Mendelian randomization mixed-scale treatment effect robust identification approach relies on (i) an assumption that the treatment effect does not vary with the possibly invalid IV on the additive scale; (ii) that the confounding bias does not vary with the possibly invalid IV on the odds ratio scale; and (iii) that the residual variance for the outcome is heteroskedastic with respect to the possibly invalid IV. In order to incorporate multiple, possibly correlated, and weak invalid IVs, a common challenge in MR studies, we develop a MAny Weak Invalid Instruments (MR MaWII MiSTERI) approach for strengthened identification and improved estimation accuracy.

91. CONTRIBUTED PAPERS: BAYESIAN METHODS FOR HIGH DIMENSIONAL/CLUSTERED DATA

Chair: Anna Heath, The Hospital for Sick Children

8:30-8:45 AM

Bayesian Analysis of Finited Mixture Model for Spherical Data

Siddhesh Kulkarni, Bristol Myers Squibb

In this project, we perform Bayesian Analysis for the von Mises Fisher (vMF) distribution on the sphere which is a common and important distribution used for directional data. We propose a new conjugate prior for the mean vector and concentration parameter of the vMF distribution. Further we prove its properties like finiteness, unimodality, and provide interpretations of its hyperparameters. Furthermore, we apply the developed methodology to Diffusion Tensor Imaging (DTI) data for clustering to explore voxel connectivity in human brain.

8:45-9:00 AM

Bayesian Ultrahigh Dimensional Variable Selection for Multivariate Mixed-Type Responses

Hsin-Hsiung Huang, University of Central Florida

It is important to develop a variable selection method for q mixed-type responses including continuous, binary and count in high-dimensional data and investigate whether the proposed method can consistently estimate the model parameters. To this end, shrinkage priors are useful for identifying relevant signals in high-dimensional data. We develop a multivariate Bayesian model with shrinkage priors model to mixed-type response generalized linear models, and we consider a latent multivariate linear regression model associated with the observable mixed-type response vector through its link function. We show that the proposed model achieves strong posterior consistency when p grows at a subexponential rate with sample size n. Furthermore, we quantify the posterior contraction rate at which the posterior shrinks around the true regression coefficients and allow the dimension of the responses q to grow as n grows. To address the non-conjugacy concern, we propose an adaptive sampling algorithm via a Polya-gamma data augmentation scheme for parameter estimation. We provide simulation studies and real-world cancer gene expression data and chronic kidney disease examples.

9:00-9:15 AM

Bayesian High-Dimensional Model Selection Consistency Under a General Prior

Min Hua, The National Cancer Institute

Bayesian model selection has become a popular approach in the high-dimensional regression analysis. The Bayesian approach selects the model which maximizes the posterior model probability as the true model. Despite the wide range of applications, the consistency of the posterior model probability is not well understood and supported in a theoretical sense. Under the Bayesian framework, the prior of model parameters plays an important role in deriving the posterior model probability. In this study, we investigate the asymptotic behavior of the posterior model probability under a general class of priors. In our high-dimensional regression setting, we allow the number of potential predictor variables to increase with the sample size. We propose sufficient conditions for the priors of model parameters to achieve posterior model probability consistency. The sufficient conditions provide useful guidelines for the specifications of priors and the corresponding hyper parameters. Our simulation and real data studies demonstrate that the sufficient conditions are crucial to the success of Bayesian model selection in the high-dimensional regression.

9:15-9:30 AM

Regularized Regression Integrating Prior Information for Classification Problems

Jingxuan He, University of Southern California

Penalized regression is a common approach for feature selection. To enhance model prediction and interpretation, some methods exist to integrate prior data during the modeling process rather than post-hoc analysis for regression. To this end, we developed an approach that implements prior-informed penalized regression for classification problems. Specifically, regression coefficients are regularized by feature-specific penalty parameters which are modeled as a log-linear function of prior covariates. Penalty vectors are estimated by empirical Bayes method instead of cross-validation and a partial quadratic approximation is implemented for an analytical solution reducing the computational complexity for multiclass outcomes. The resulting marginal likelihood is optimized by a designed iterative reweighted-L2 algorithm. Through simulation studies and applied examples, we demonstrate our method's improved prediction accuracy, feature selection, and effect estimation compared with regular penalized models. We discuss the relationship to relevant vector machine and present software application in R package.

9:30-9:45 AM

A Non-Parametric Factor Model with Variable Selection Under Non-Normal Distribution

Yanan Zhang, University of South Carolina, Department of Epidemiology and Biostatistics

Variable selection is often used to identify important variables and improve prediction performance. For correlated data, in available variable selection methods, some of them either do not work well due to violation of underlying independent assumptions or do not provide information on correlation patterns. Factor analysis can explore the latent structure by grouping the correlated covariates into independent factors but lacks the flexibility to accommodate the nonparametric settings. In this article, we propose a Bayesian nonparametric model in the framework of factor analysis. The developed model relaxes the normality assumption of the linear regression, allowing for grouping the correlated covariates and selection of the latent factors simultaneously. The performance of the proposed method is evaluated based on comprehensive simulation studies, where the four commonly used methods are compared including LASSO regression, grouped LASSO regression, factor analysis, and sparse Bayesian infinite factor model. In addition, a real dataset from the genetic vs environment in the scleroderma outcomes study (GENISOS) cohort is analyzed using the proposed method with comparisons.

9:45-10:00 AM

Bayesian Hierarchical Models for Mitochondrial DNA Data

Jenny Brynjarsdottir, Case Western Reserve University

We discuss Bayesian hierarchical modeling (BHM) approaches to model and interpret DNA and RNA sequence data from the human mitochondrial genome (mtDNA and mtRNA). Many human diseases are tied to mitochondrial dysfunction. Complicating the analysis of the mitochondrial genomic contributions to these diseases is the fact that mtDNA mutational content is extremely fluid from organ to organ within the same individual, and even within the same organ over time. Since the phenotypic effect of a mitochondrial variant depends on its heteroplasmy level (multiple distinct mtDNA haplotypes witin an individual, tissue, or cell), understanding the dynamics underlying the ebb and flow of mtDNA variants over time, within different tissues, and across generations is necessary to address many aspects of human health. We propose a BHM framework that can incorporate various types of count data that can capture these dynamics.

10:00-10:15 AM

Pre-Trained Knowledge Guided Transfer Learning for Identifying Clinical Subphenotypes of Multisystem Inflammatory Syndrome in Children Using National Pediatric Recover Cohort

Xiaokang Liu, University of Pennsylvania

As one of the most severe post-acute sequelae of SARS-CoV-2 infection in children, multisystem inflammatory syndrome in children (MIS-C) involves serious inflammation and has highly heterogeneous manifestations. Disentangling complex manifestations of MIS-C by finding its subphenotypes helps identify children at high risk of severe outcomes and determine more targeted therapies. Identifying subphenotypes usually requires large sample sizes and may benefit from combing data from multiple hospitals. However, joint analysis often faces two challenges: prohibition of sharing patient-level data due to privacy concerns and heterogeneity in population across sites. We propose a one-shot aggregated data-based method to transfer knowledge of the shared subphenotypes pre-trained on one site to another site to achieve joint learning. Site-specific subphenotype mixing proportions are used to explain between-site heterogeneity. Simulation studies show that the resultant estimator performs close to the pooled data-based estimator. We use this method to transfer the MIS-C subphenotypes learned using electronic health records from PEDSnet to an external site to achieve joint analysis.

92. CONTRIBUTED PAPERS: EFFICIENT/ROBUST ESTIMATION UNDER A CAUSAL INFERENCE FRAMEWORK

Chair: Siyu Heng, New York University

8:30-8:45 AM

Comparing Vaccine Immunogenicity Across Trials with Different Populations and Study Designs

Yutong Jin, Emory University

Recent studies has revealed that effective vaccines contributed tremendously to the prevention of infectious diseases. The effectiveness of vaccine is typically measured in randomized efficacy trials using clinical endpoints. However, this process is often time-consuming and has significantly slowed the vaccine research. To reduce the development time, one promising solution is to assess surrogate endpoints, like immune responses, that are predictive of vaccine efficacy. However, the measurement of such immune responses is expensive so that a two-stage sampling strategy is employed for large trials. Additionally, trials of difference vaccines may be conducted in diverse study populations. It is thus hard to provide an objective comparison of vaccine immunogenicity across different vaccines directly. To address this issue, we propose a framework to identify appropriate causal estimands and estimators that can be used to provide standardized comparisons of vaccine immunogenicity across trials. Our estimators are well performed and enjoy robustness properties. This method is then applied to data from four recent HIV vaccine trials.

8:45-9:00 AM

Efficient Nonparametric Estimation of Incremental Propensity Score Effects with Clustered Interference

Chanhwa Lee, Department of Biostatistics, University of North Carolina at Chapel Hill

Interference occurs when a unit's treatment affects another unit's outcome. In some cases, units may be grouped into clusters such that it is reasonable to assume no interference between units in different clusters, i.e., there is clustered interference. Various causal estimands have been proposed to quantify treatment effects under clustered interference in observational data, but these estimands either entail treatment policies of little real-world relevance or are based on parametric propensity score models. Here, we propose new causal estimands based on incremental changes to propensity scores which may be more relevant in many contexts and are not based on parametric models. Nonparametric sample splitting estimators of the new estimands are constructed, which allow for flexible data-adaptive estimation of nuisance functions and are consistent, asymptotically normal, and efficient, converging at the usual parametric rate. Simulations show the finite sample performance of the proposed estimators. The proposed methods are applied to evaluate the effect of water, sanitation, and hygiene facilities on the diarrhea incidence among children in Senegal.

9:00-9:15 AM

Identifying and Estimating Effects of Sustained Interventions Under Parallel Trends Assumptions

Audrey Renson, New York University Grossman School of Medicine

Existing methods to estimate effects of sustained interventions typically require no unmeasured confounding, which can be questionable in observational studies. Differences-in-differences relies instead on the parallel trends assumption, allowing for some types of unmeasured confounding, but most existing difference-in-differences implementations are limited to point treatments in restricted subpopulations. We derive identification results for population effects of sustained treatments under parallel trends. In particular, if all individuals begin follow-up with exposure status consistent with the treatment plan of interest but may later deviate, a version of Robins' g-formula identifies the intervention-specific mean under SUTVA, positivity, and parallel trends. We develop consistent asymptotically normal estimators based on inverse-probability weighting, g-computation, and a double robust estimator that combines both. Simulation studies support the use of the proposed estimators at realistic sample sizes. The methods are used to estimate effects of a hypothetical federal stay-at-home order on all-cause mortality during the COVID-19 pandemic in spring 2020 in the United States.

9:15-9:30 AM

Application of Marginalized Zero-Inflated Models when Mediators Have Excess Zeroes

Andrew Sims, University of Alabama at Birmingham Department of Biostatistics

Mediation analysis has recently become more popular as researchers are interested in establishing mechanistic pathways for intervention. Although available methods have increased, there are limited options for mediation analysis with zero-inflated count variables. Current methods do not obtain population average effects of mediation effects, the quantity most desired in applied research. In this paper we propose an extension of the counterfactual approach to mediation to scenarios where the mediator is a count variable with excess zeroes by utilizing the Marginalized Zero-Inflated Poisson Model (MZIP) for the mediator model. We derive direct and indirect effects for continuous, binary, and count outcomes, as well as adapt to allow mediator-exposure interactions. Our proposed work allows straightforward calculation of direct and indirect effects for the overall population mean values of the mediator. We apply this novel methodology to an application observing how alcohol consumption may explain sex differences in cholesterol and assess model performance via a simulation study comparing the proposed MZIP mediator framework to Poisson and linear mediator regression models.

9:30-9:45 AM

Doubly Robust Estimation and Sensitivity Analysis for Marginal Structural Quantile Models

Chao Cheng*, Yale University

The marginal structure quantile model (MSQM) is a useful tool to characterize the causal effect of a time-varying treatment on the full distribution of potential outcomes. To date, only the inverse probability weighting (IPW) approach has been developed to identify the causal parameters in MSQM, which requires correct specification of the propensity score models for the treatment assignment mechanism. We propose a doubly robust approach for the MSQM under the semiparametric framework. The proposed approach is consistent if either of the models for treatment assignment or the models for potential outcome distribution are correctly specified, and is locally efficient if both models are correct. In addition, we develop a new sensitivity analysis strategy to investigate the robustness of MSQM estimators when the sequential ignorability assumption is violated. We apply the proposed methods to the Yale New Haven Health System Electronic Health Record data to study the effect of antihypertensive medications to patients with severe hypertension and assess the robustness of findings to unmeasured confounding.

9:45-10:00 AM

Double Sampling for Informatively Missing Data in Electronic Health Record-Based Comparative Effectiveness Research

Alexander Levis, Carnegie Mellon University

Missing data arise almost ubiquitously in applied settings. When data are missing not at random (MNAR) with respect to measured covariates, sensitivity analyses are often proposed as a post-hoc solution, though these may not always yield concrete conclusions. Motivated by an electronic health records-based study of long-term outcomes following bariatric surgery, we consider the use of double sampling as a means to mitigate MNAR outcome data when the statistical goals are estimation and inference regarding causal effects. We describe assumptions under which the joint distribution of confounders, treatment, and outcome is identified under this design, and derive efficient and robust estimators of the average causal treatment effect under a nonparametric model and under a model assuming outcomes were initially missing at random. We compare these in simulations to an approach that adaptively estimates based on evidence of violation of the missing at random assumption. We also show that the proposed double sampling design can be extended to handle arbitrary coarsening mechanisms, and derive nonparametric efficient estimators of any smooth full data functional.

93. CONTRIBUTED PAPERS: LATENT VARIABLES

Chair: Yan Li, The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences

8:30-8:45 AM

Node-Wise Community Detection in Network Data with the Presence of Interaction-wise Latent Topic Labels

Yuhua Zhang, University of Michigan

Scientists are increasingly interested in discovering community structure from modern relational data arising on large-scale social networks. While many methods have been proposed for learning community structure, few account for the fact that interactions may exhibit different topics. In this paper, we introduce a novel method that integrates the interaction-wise latent topic labels into node-level community detection. In particular, our method allows the incorporation of prior knowledge from interaction-wise textual information. A computationally tractable variational inference algorithm is derived. We demonstrate the effectiveness of our method in various simulation settings. We apply the method to TalkLife, a large-scale online peer-to-peer support network.

8:45-9:00 AM

Latent Variables Reveal Structure in Multiple Outcomes Models for Developmental Endpoints

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 effects of prenatal methylmercury exposure on childhood neurodevelopment. Rather than cluster individual subjects, we cluster test outcomes within a multiple outcomes model to improve estimation of the exposure effect and the model fit diagnostics. By using information on exposures in the data to nest the outcomes into groups called domains, the model more accurately reflects the shared characteristics of neurodevelopmental outcomes. The paradigm allows for sampling from the posterior distribution of the grouping parameters; thus, inference can be made on group membership and their defining characteristics. We avoid the often difficult requirement of a priori identification of the total number of groups by incorporating a Dirichlet process prior. In doing so, we estimate exposure effects on neurodevelopment while shrinking effects within and between the domains selected by the data.

9:00-9:15 AM

Latent Transition Analysis to Classify the Food Environment

Kelsey Alexovitz, Drexel University

Examining how health outcomes are influenced by the food environment (FE) is challenging given that multiple features (e.g. different types of food outlets) can be used to measure the FE, and that the features vary in spatial proximity to subjects and over time. Moreover, measures of food environment features have an excess of zeros. We propose a latent transition analysis (LTA) model to classify the food environment that uses zero-inflated Poisson distributions to model the observed multivariate count data. Using simulations, we examine how our approach compares to classical LTA that uses multinomial distributions for observed variables. Simulations show that zero-inflated Poisson distributions for observed variables perform the best when there are more than 2 latent classes. We apply our modeling strategy to classify the food environment near California public schools. The derived classification of the food environment near schools can be used to inform interventions and other preventative strategies.

9:15-9:30 AM

A Semi-Parametric Model for Dispersed Count Data

Cornelis Potgieter, Texas Christian University

We consider a model where bounded multivariate count data serve as a proxy for a continuous latent trait. Each observed count is the number of successes in a fixed series of trials. However, the data may exhibit under- or over-dispersed relative to the binomial model. For model estimation, we propose a semiparametric framework. The latent trait is assumed to follow a standard normal distribution, but the count variables are only defined through their first two conditional moments with respect to the latent trait. The count variables are further assumed to adhere to a group (testlet) structure. A generalized method of moment (GMM) estimator of the model parameters is defined. Asymptotic normality of this estimator is established and finite-sample properties are explored in a simulation study. The methodology is illustrated in oral reading fluency assessment data for elementary school children. In the sample, each child was given a number of passages to read aloud and the number of words read correctly per sentence were recorded. We apply our semiparametric method to this dataset and also illustrate recovery of latent reading ability.

9:30-9:45 AM

ARTdeConv: A R Package for the Flexible Adaptive Regularized Tri-Factor Nonnegative Matrix Factorization Method for Deconvolution of Bulk Tissue Cell Types

Tianyi Liu, University of North Carolina at Chapel Hill

Cell type deconvolution of bulk gene expression profiles is essential for downstream data analysis and characterizing disease states. However, many deconvolution methods either require complete cellular gene expression signatures or rely entirely on unsupervised methods, not utilizing any biological information. Furthermore, existing methods often neglect the varying sequencing platforms where bulk data and gene signatures are generated and cell sizes across cell types, leading to biased proportion estimates. We propose an Adaptive Regularized Tri-factor nonnegative matrix factorization method for deconvolution (ARTdeConv) to address the aforementioned issues. We prove convergence properties of the algorithm, and further demonstrate its accuracy in deconvolution. Namely, we compare to existing methods using simulations and on real-world bulk sample data from studies in chlamydia and H1N1 influenza that involve human whole blood samples, where gene signatures of neutrophils are unobtainable.

9:45-10:00 AM

Latent Class Proportional Hazards Regression with Heterogeneous Survival Data

Teng Fei, Memorial Sloan Kettering Cancer Center

Heterogeneous survival data are commonly present in chronic disease studies. Delineating meaningful disease subtypes directly linked to the survival outcome can generate useful scientific implications. In this work, we develop a latent class proportional hazards (PH) regression framework to address such an interest. We propose mixture proportional hazards modeling, which flexibly accommodates class-specific covariate effects while allowing for the baseline hazard function to vary across latent classes. Adapting the strategy of nonparametric maximum likelihood estimation, we derive an Expectation-Maximization (E-M) algorithm to estimate the proposed model. Extensive simulation studies are conducted, demonstrating satisfactory finite-sample performance of the proposed method as well as the predictive benefit from accounting for the heterogeneity across latent classes. We further illustrate the practical utility of the proposed method through an application to a mild cognitive impairment (MCI) cohort in the Uniform Data Set.

10:00-10:15 AM

Joint Sparse Factor Regression Models for Analyzing High-Dimensional Data

Alexander Quinter, University of North Carolina at Chapel Hill

Factor analysis provides an approach for dimension reduction when working with big data. It allows researchers to represent an extensive number of correlated variables via a set of latent factors. Traditional methods lack properties desirable for analyzing big data. Factor regression models in the literature assume the same set of latent factors underlie both the independent and dependent variables. This assumption is especially restrictive when it is believed the latent factors associated with dependent variables are a function of the latent factors associated with independent variables. The lack of an approach that can address these issues requires the development of a new method for joint sparse factor regression models. We propose such a method here and substantiate it via simulation studies that indicate the method can correctly identify latent factors underlying the independent and dependent sets of variables and accurately estimate the entries of the factor loading matrix estimates. We apply our method to the COVIDiSTRESS dataset to demonstrate our model has superior performance to classical models, while also identifying the latent constructs intrinsic to the data.

94. CONTRIBUTED PAPERS:INDIVIDUALIZED TREATMENT RULES/RISK PREDICTION

Chair: Brian Egleston, Fox Chase Cancer Center

8:30-8:45 AM

Multi-Objective Tree-Based Reinforcement Learning for Estimating Tolerant Dynamic Treatment Regimes

Yao Song*, University of Michigan, Ann Arbor

A dynamic treatment regime (DTR) is a sequence of decision rules that adapt to the time-varying states of an individual. It provides a vehicle for operationalizing a clinical decision support system. However, many real-world problems involve multiple competing priorities, and decision rules differ when trade-offs are present. We propose a concept of a tolerant regime that gives a set of individualized feasible decision rules under a pre-specified tolerance rate. A multi-objective tree-based reinforcement learning (MOT-RL) method is presented to estimate the tolerant DTR (tDTR) that optimizes multiple objectives in a multi-stage setting. At each stage, MOT-RL constructs an unsupervised decision tree by modeling the counterfactual mean outcome of each objective via semiparametric regression and maximizing a purity measure constructed by the scalarized augmented inverse probability weighted estimators (SAIPWE). MOT-RL is robust, efficient, easy to interpret, and flexible across different problem settings. With the proposed method, we identify two-stage chemotherapy regimes that simultaneously reduce disease burden and prolong the survival of advanced prostate cancer patients.

8:45-9:00 AM

A Bayesian Decision Framework for Optimizing Sequential Combination Antiretroviral Therapy in People with HIV

Wei Jin, Department of Applied Mathematics and Statistics, Johns Hopkins University

Numerous adverse effects have been reported for combination antiretroviral therapy (cART) despite its remarkable success on viral suppression in people with HIV (PWH). To improve long-term health outcomes for PWH, there is an urgent need to design personalized cART with the lowest risk of comorbidity in the field of precision medicine for HIV. Large-scale HIV studies offer us opportunities to optimize cART in a data-driven manner. However, the large number of possible drug combinations for cART makes the estimation a high-dimensional problem, imposing challenges in both statistical inference and decision-making. We develop a two-step Bayesian decision framework for optimizing sequential cARTs. In the first step, we model individuals? longitudinal observations using a multivariate Gaussian process. In the second step, we build a probabilistic generative model for cARTs and design an uncertainty-penalized policy optimization. Applying the method to the Women?s Interagency HIV Study, we demonstrate its clinical utility in assisting physicians to make effective treatment decisions, serving the purpose of both viral suppression and comorbidity risk reduction.

9:00-9:15 AM

Comparing Machine Learning Methods for Estimating Individual Treatment Effects by Combining Data from Multiple Randomized Controlled Trials

Carly Lupton-Smith, Johns Hopkins Bloomberg School of Public Health Department of Biostatistics

Individualized treatment decisions can improve efficiency in medical practice, but using data to make these decisions in a reliable, powerful and generalizable way is difficult with a single dataset. Leveraging multiple randomized controlled trials allows for the combination of datasets with unconfounded treatment assignment to improve the power to estimate treatment effect heterogeneity at the individual level. In this paper, we discuss several non-parametric approaches for estimating treatment effect heterogeneity using data from multiple trials. We compare different single-study methods and different ways of aggregating those to the multi-trial setting through a simulation study, with data generation scenarios that have varying levels of cross-trial heterogeneity. We find that methods that directly allow for heterogeneity of the treatment effect across trials perform better than methods that do not, and that the choice of single-study method matters based on the complexity of the treatment effect. We discuss which methods perform well in each setting and then apply them to three randomized controlled trials comparing the effects of drugs for major depressive disorder.

9:15-9:30 AM

Individualized Treatment Rules and treatment effects estimation Following Multiple Imputation

Jenny Shen, University of Pennsylvania

Data-driven optimal treatment strategies can benefit patients, care providers, and others by improving clinical outcomes and lowering healthcare costs. A treatment decision rule is a function that maps patient-level information to a recommended treatment. An optimal treatment decision rule maximizes a population-level distributional summary. However, guidance for estimating and evaluating optimal treatment decision rules in the presence of missing data is limited. Our work is motivated by the Social Incentives to Encourage Physical Activity and Understand Predictors (STEP UP) study, a multi-interventional randomized trial for physical activity. Study participants were given wearable devices to record daily step counts, a measure subject to missingness. In the primary analysis, multiple imputation (MI) was used. We describe two frameworks for estimation and evaluation of an optimal treatment decision rule following MI and compare their performance using simulated data. We then provide an illustrative analysis to estimate and evaluate an optimal decision rule from the STEP UP data with a focus on practical considerations such as choosing an appropriate number of imputations.

9:30-9:45 AM

Dynamic Treatment Regime Characterization via a Value Function Surrogate with an Application to Partial Compliance

Nikki Freeman, University of North Carolina at Chapel Hill

Motivated by the gap the between precision medicine evidence and clinical decision support, we bring a new perspective to the dynamic treatment regime (DTRs) learn- ing problem by focusing on tools for understanding DTRs and understanding DTRs in real world contexts. Through an application of the precision medicine framework to partial compliance to wound management in peripheral artery disease, we demonstrate the utility of a Gaussian process surrogate for value function approximation for finite parametric classes of DTRs. We use the Gaussian process surrogate to show the feasibility of learning an optimal DTR through Bayesian optimization and characterizing an entire class of policies.

9:45-10:00 AM

Bayesian Estimation of a Semi-Parametric Recurrent Event Model of Multiple Cancer Types for Personalized Risk Prediction in Cancer Survivors

Hoai Nam Nguyen, Department of Statistics, Rice University

The number of cancer survivors in the US is expected to reach 20 million by 2026. Various studies have shown that the risk of subsequent primary depends on the characteristics of the first. To quantify cancer-specific risks beyond the first primary, we propose a Bayesian semiparametric framework, where the occurrences of each cancer type follow a non-homogeneous Poisson process that depends on the type and timing of the first primary. Upon applying our model to a Li-Fraumeni Syndrome (LFS) dataset, we contribute the first penetrance estimates for lung and prostate cancers, which produce accurate risk predictions on an independent dataset. For family-based datasets with partially observed genotypes, we use the peeling algorithm to address the inheritance of TP53 mutations, and the ascertainment-corrected joint approach to correct for ascertainment bias. Validation on an independent dataset shows the promising performance of the model when making cancer-specific predictions of the second primary. This performance is far above a na?ve model that assumes independence between the first and second primary, indicating the usefulness of our models in real clinical settings.

95. CONTRIBUTED PAPERS: STATISTICAL METHODS FOR HIGH-THROUGHPUT SEQUENCING DATA

Chair: Sarah Lotspeich, Wake Forest University

8:30-8:45 AM

Spatial Autocorrelation and Implications on Differential Expression Analysis Involving Spatial Transcriptomics

Brooke Fridley, Moffitt Cancer Center

Recent years have seen an increased interest in spatially resolved transcriptomics (ST). Often researchers conduct differential expression analysis on tissue niches. Currently, most methods for determining differential expressed (DE) genes use models that do not account for the spatial dependency between the measured locations. Hence, using multiple datasets from three ST technologies (Visium, GeoMx, CosMx) we compared DE analysis results between models accounting for or not the spatial dependency. For the spatial modeling, mixed models were fit with a spherical error model. For GeoMx, we found that regions of interest often show low spatial autocorrelation, with the spatial model being significantly better than the non-spatial model in roughly 5% of genes. However, in the case of Visium, where tissues are more densely sampled, the spatial models often outperformed the non-spatial model in 10-40% of genes. Thus, we recommend that researchers should consider fitting spatial mixed models when determining DE genes and at a minimum, complete a sensitivity analysis for top DE genes using a spatial model to confirm the results.

8:45-9:00 AM

Statistical Methods for Microbiome Driven Integration of Omics Data

Rebecca Deek, University of Pennsylvania

Advances in technology have contributed to an increasing number of studies involving sequencing of the genome, transcriptome, proteome, metabolome, or microbiome. Declining sequencing costs have enabled large-scale sequencing of two or more of the “omes” in a single study. Microbial sequencing studies in particular are moving towards understanding the functional capacity of the microbiome, often described in terms of metabolomics and proteomics. Despite this, there remains a gap in understanding if available methodology and software are well-suited for such data. Our work reviews commonly used existing statistical tools for identifying global and feature-specific associations between the microbiome and other host omics datasets. We also discuss methods for incorporating clinical factors such as treatment and disease status or progression. Finally, we discuss best practices and where new methodologies are needed.

9:00-9:15 AM

Quantifying the HIV Reservoir with Dilution Assays and Deep Viral Sequencing

Brian Richardson, Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina at Chapel Hill

The quantitative viral outgrowth assay (QVOA) is commonly used to measure the latent HIV reservoir in individuals living with HIV on antiretroviral treatment. Eliminating this reservoir is the goal of HIV cure research. Data from QVOA can be used to estimate the reservoir size, i.e., the infectious units per million (IUPM) of HIV persistent resting CD4+ cells. A variation of QVOA, the Ultra Deep Sequencing Assay of the Outgrowth Virus (UDSA), was recently developed that quantifies the number of viral lineages within infected cells. This additional sequencing information can potentially be used to improve IUPM estimation. Maximum likelihood has been used to estimate the IUPM from QVOA and UDSA data, but existing approaches (i) assume all assay wells have the same number of cells and (ii) yield upwardly biased estimates. We propose an extension of these methods which accommodates assays with wells sequenced at multiple dilution levels and includes a bias-corrected estimator of the IUPM. The accuracy and efficiency of the proposed methods are demonstrated through simulations. The proposed methods are applied to data from the University of North Carolina HIV Cure Center.

9:15-9:30 AM

ISLET: Individual Reference Panel Recovery Improves Cell-Type-Specific Inference

Hao Feng, Case Western Reserve University

We propose a statistical framework ISLET to infer individual-specific and cell-type-specific transcriptome reference panels. ISLET rigorously models the repeatedly measured bulk gene expression data, to optimize the usage of shared information within the same subject. ISLET is the first available method to achieve individual-specific reference estimation. Using simulation studies, we show favorable performance of ISLET in the reference estimation and downstream cell-type-specific differentially expressed gene testing. We applied ISLET on longitudinally profiled transcriptomes in blood samples from a large observational study of young children, and confirmed the cell type-related gene signatures for pancreatic islet autoantibody. ISLET is available at <a href="https://bioconductor.org/packages/ISLET">https://bioconductor.org/packages/ISLET</a>.

9:30-9:45 AM

Flexible Semiparametric Methods for Differential Abundance Analysis in Microbiome Studies

Olivier Thas, Hasselt University

In many microbiome studies one is interested in testing for differential abundance between two or more conditions. Many methods have been proposed in recent years and many of them have been empirically evaluated in simulation studies. So far the conclusion is that methods based on distributional assumptions (e.g. the negative binomial) may perform worse than others when these assumptions are violated. In recent studies, particularly the ANCOM-BC method seems to perform well.

In this talk, we present two new semiparametric methods for testing for differential abundance. A first method is based on probabilistic index models (PIMs), and a second method is developed along the lines of ANCOM-BC, without requiring the log-transformation and allowing for a mean-variance relationship. Our methods are flexible (e.g. allow for covariate-adjustment), do not rely on stringent distributional assumptions, control the FDR and are competitive to other well performing methods. Our methods are evaluated and compared to competitors in a comprehensive simulation study.

9:45-10:00 AM

BBQ: Better Base Qualities for Next-Generation Sequencing

Wenyu Gao, Department of Biostatistics, Harvard University

Despite the ubiquitous application of massively parallel "next generation" sequencing, detailed error models describing read-level error probabilities are lacking for all but specialized applications. In this talk, we present a novel base-quality recalibration method (BBQ) with demonstrated improvements in calibration performance over current state-of-the-art methods. The BBQ method relies on training a statistical model for both PCR-based library creation and sequencing-synthesis errors using large numbers of observed errors of each type. We evaluate the performance of BBQ in terms of sensitivity and specificity for the detection of minimal residual disease in cancer recurrence monitoring.

10:00-10:15 AM

A Novel Gene-Based Test for Next Generation Sequencing Studies Based on a Bayesian Variable Selection of Rare Variants

Jingxiong Xu, Lunenfeld-Tanebaum Research Institute

A usual paradigm for detecting rare variants (RVs) associated with complex human diseases is to perform a gene-based association test. However, the inclusion of all RVs in a gene-based test might reduce its power. As an alternative, we propose a novel strategy that performs a variable selection of the RVs to be considered in the gene-based test as a powerful approach. Our Bayes Factor test statistic is based on generalized linear model and its conjugate prior, which can handle outcomes of different types (continuous and categorical), variant annotations and unbalanced designs. A birth-death MCMC algorithm is used to select the important RVs to be considered in the test. Through simulation studies, we show that including a variable selection step of RVs in a gene-based test is a more powerful approach than considering all RVs in the gene as proposed in current tests. Variant annotations can be used to define prior in our Bayesian approach. Application to UK Biobank sequencing data with lung cancer outcome show that our method leads to new gene and RV discoveries.

96. CONTRIBUTED PAPERS: MACHINE LEARNING - METHODS AND APPLICATIONS

Chair: Guanqun Cao, Auburn University

8:30-8:45 AM

Examining an Alternative Sampling Approach for Selecting the Training Dataset in the Random Forest Algorithm that Considers the Spatial Nature of the Data

Melissa Meeker, Drexel University

The random forest algorithm is often applied to spatial datasets; however, the random forest uses a simple random sample to split the data into training and test sets which ignores the dependency among spatial observations. In this paper, we compare the implementation of a geographically stratified sampling approach to the traditional simple random sample in the random forest algorithm for simulated spatial data, as well as in data describing Philadelphia Police Department pedestrian investigations. From the simulated data, we found that implementing the geographically stratified sampling technique improves predictive performance in less densely observed regions but reduces predictive performance in more densely observed regions of the spatial context. In the pedestrian investigations dataset, we found that implementing the geographically stratified sampling technique improves sensitivity in the training data but worsens sensitivity in the test data. Overall, the modification balances predictive performance across the spatial context by allowing the random forest to be trained on more observations from the less dense regions.

8:45-9:00 AM

Flexible Estimation of the Conditional Survival Function via Observable Regressions

Charles Wolock, Department of Biostatistics, University of Washington

The conditional survival function of a time-to-event outcome subject to censoring and truncation is a common estimation target in survival analysis. This parameter may be of scientific interest and often appears as a nuisance in semiparametric settings. In addition to parametric and semiparametric methods (e.g., based on the Cox model), flexible machine learning approaches have been developed to estimate the conditional survival function. However, many of these methods are targeted toward risk stratification rather than function estimation. Others apply only to discrete time settings or require inverse probability of censoring weights, which can be as difficult to estimate as the outcome survival function itself. We propose a decomposition of the conditional survival function in terms of observable regressions in which censoring and truncation play no role. This allows application of off-the-shelf learning methods rather than only approaches that explicitly handle the complexities of survival data. We outline estimation procedures based on this decomposition, assess their performance via numerical simulations, and demonstrate their use on data from an HIV vaccine trial.

9:00-9:15 AM

Cellcano: Supervised Cell Type Identification for Single Cell ATAC-Seq Data

Wenjing Ma, Department of Computer Science, Emory University

Computational cell type identification (celltyping) is a fundamental step in single-cell omics data analysis. Supervised celltyping methods have gained increasing popularity in single-cell RNA-seq data because of the superior performance and the availability of high-quality reference datasets. Recent technological advances in profiling chromatin accessibility at single-cell resolution (scATAC-seq) have brought new insights to the understanding of epigenetic heterogeneity. With continuous accumulation of scATAC-seq datasets, supervised celltyping method specifically designed for scATAC-seq is in urgent need. In this work, we develop Cellcano, a novel computational method based on a two-round supervised learning algorithm to identify cell types from scATAC-seq data. The method alleviates the distributional shift between reference and target data and improves the prediction performance. We systematically benchmark Cellcano on 50 well-designed experiments from various datasets and show that Cellcano is accurate, robust, and computational efficient.

9:15-9:30 AM

Transcriptomic Congruence and Selection of Representative Cancer Models Towards Precision Medicine

Jian Zou, Department of Biostatistics, University of Pittsburgh

Cancer models are instrumental to substitute for human studies and expedite basic, translational and clinical cancer research. For a given cancer subtype, a wide selection of models, such as cell lines, patient-derived xenografts, tumoroids and genetically modified murine models, are often available to researchers. However, how to quantify their congruence to human tumors and to select the most appropriate cancer model is a largely unsolved issue. Here, we develop Congruence Analysis and Selection of CAncer Models (CASCAM), a statistical and machine learning framework for authenticating and selecting the most representative cancer models in pathway-specific and drug-relevant context using transcriptomic data. CASCAM offers harmonization between tumor and cancer model omics data, interpretable machine learning for congruence quantification, mechanistic investigation, and pathwaybased topological visualization to determine the final cancer model selection. The workflow is presented using breast cancer invasive lobular carcinoma (ILC) subtype, while the method is generalizable to any cancer subtype for precision medicine development.

9:30-9:45 AM

Data-Driven Method for Combining Measurements from a Group of Heterogeneous Raters for the Evaluation of a New Device

Qi Yu, Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University

With advancements in technology, many computer-aided diagnostic devices (CAD) are introduced to detect complex diseases. This work is motivated by the need to develop a framework to evaluate the performance of a newly developed CAD. When a gold standard is absent, often in practice, the evaluation of a new CAD is performed by comparing its measurement to the average of multiple opinions from clinicians regarded as the best available standard. while this approach may lead to biased evaluations as the clinicians may have different experiences and accuracy levels. In this work, we propose a novel weighting strategy to combine measurements from a heterogeneous group of clinicians. Specifically, an unsupervised induction method is proposed to assign higher weights to clinicians who consistently agree with others and to give lower weights to those who mostly disagree with others, providing a fair evaluation of a new device according to the consistent opinions among clinicians. Our method is compatible with any existing agreement measures. We demonstrate the practical utility of the proposed method via extensive simulation studies and an application to renal study data.

9:45-10:00 AM

Optimizing Contingency Management in Substance Use Disorder Treatment Using Off-Policy Policy Evaluation

Younggeun Kim, Adjunct Associate Research Scientist

Contingency Management (CM), a policy of providing financial incentives to subjects who abstain from substance use to promote treatment effects, has played an important role in various disorders including substance use disorders. Maximizing the treatment effects for given budget constraints is of great interest in CM. However, collecting data every time to evaluate each CM policy is cost[1]expensive and takes a long time. To overcome this challenge, we apply off-policy policy evaluation, a reinforcement learning technique to evaluate the new policy with a dataset collected from behavior policy. Our study integrates various CM study datasets, evaluates various CM policies, and examines a long-standing hypothesis in the field of CM: The treatment is agnostic to the substance of abuse?

Wednesday, March 22, 2023 | 10:30-12:15 PM

97. TENSOR-BASED METHODS IN BIOMEDICAL DATA SCIENCE

Organizer: Lei Liu, Washington University in St. Louis
Chair: Lili Liu, Washington University in St. Louis

10:30-10:55 AM

Functional Tensor Singular Value Decomposition with Applications in High-order Longitudinal Data Analysis

Anru Zhang, Duke University

In this talk, we introduce the functional tensor singular value decomposition (FTSVD), a novel dimension reduction framework for tensors with one functional mode and several tabular modes. The problem is motivated by high-order longitudinal data analysis. Our model assumes the observed data to be a random realization of an approximate CP low-rank functional tensor measured on a discrete time grid. Incorporating tensor algebra and the theory of Reproducing Kernel Hilbert Space (RKHS), we propose a novel RKHS-based constrained power iteration with spectral initialization. Our method can successfully estimate both singular vectors and functions of the low-rank structure in the observed data. With mild assumptions, we establish the non-asymptotic contractive error bounds for the proposed algorithm. The superiority of the proposed framework is demonstrated in the analysis of real longitudinal microbiome data.

10:55-11:20 AM

Tensor in Multivariate Categorical Response Regression

Xin Zhang, Florida State University

In many modern regression applications, the response consists of multiple categorical random variables whose probability mass is a (tensor) function of a common set of predictors. We will discuss tensor-based modeling strategies in such regression problems of multiple categorical responses on high-dimensional predictors. We propose a new mixture of regressions model that can be fitted using an efficient and scalable penalized expectation-maximization algorithm. Low-rank tensor structures will be used for intrinsic dimension reduction and variable selection. We demonstrate the encouraging performance of our method through both simulation studies and an application to modeling the functional classes of genes.

11:20-11:45 AM

Multivariate Temporal Point Process Regression

Xiwei Tang, University of Virginia

Point process modeling is gaining increasing attention, as point process type data are emerging in a large variety of scientific applications. In this article, motivated by a neuronal spike trains study, we propose a novel point process regression model, where both the response and the predictor can be a high-dimensional point process. We model the predictor effects through the conditional intensities using a set of basis transferring functions in a convolutional fashion. We organize the corresponding transferring coefficients in the form of a three-way tensor, then impose the low-rank, sparsity, and subgroup structures on this coefficient tensor. These structures help reduce the dimensionality, integrate information across different individual processes, and facilitate the interpretation. We develop a highly scalable optimization algorithm for parameter estimation. We derive the large sample error bound for the recovered coefficient tensor, and establish the subgroup identification consistency, while allowing the dimension of the multivariate point process to diverge. We demonstrate the efficacy of our method through both simulations and a cross-area neuronal spike trains analysis in a sensory cortex study.

11:45-12:10 PM

High-Order Joint Embedding for Multi-Level Link Prediction

Yubai Yuan, The Pennsylvania State University

Link prediction infers potential links from observed networks, and is one of the essential problems in network analyses. In contrast to traditional graph representation modeling which only predicts two-way pairwise relations, we propose a novel tensor-based joint network embedding approach on simultaneously encoding pairwise links and hyperlinks onto a latent space, which captures the dependency between pairwise and multi-way links in inferring potential unobserved hyperlinks. The major advantage of the proposed embedding procedure is that it incorporates both the pairwise relationships and subgroup-wise structure among nodes to capture richer network information. In addition, the proposed method introduces a hierarchical dependency among links to infer potential hyperlinks, and leads to better link prediction. In theory we establish the estimation consistency for the proposed embedding approach, and provide a faster convergence rate compared to link prediction utilizing pairwise links or hyperlinks only.

98. METHODS FOR INTEGRATING AND ANALYZING LARGE HETEROGENEOUS MICROBIOME DATA SETS

Organizer: Ekaterina Smirnova, Virginia Commonwealth University
Chair: Lucia Tabacu, Old Dominion University

10:30-10:55 AM

Integrative Causal Analysis of Microbiome and Metabolomics Data in Longitudinal Microbiome Studies

Hongzhe Li, University of Pennsylvania

Longitudinal microbiome studies aim to understand how a treatment or an environmental exposure such as diet shapes gut microbiome, leading to changes of host metabolomics and clinical outcomed through production, degradation or modification of metabolites. While most current methods focus on associations or correlations between microbiome and metabolomics, these longitudinal data provide us the possibility of understanding possible causal relationships among microbiome, metabolomics and clinical outcomes. I will present methods of mediation analysis and graohical model to identify possible causal taxa and their associated metabolites that may lead to clinical outcome. I will show results of our analysis of two ongoing microbiome studies at the University of Pennsylvania, including the Infant Growth and Microbiome Study (IGram) and a randomized trial comparing the specific carbohydrate diet to a mediterranean diet in adults with Crohn's disease (DINE-CD).

10:55-11:20 AM

Assessing the Conditional Correlation Between Individual Genomic Features and Microbial Taxa

Michael Wu, Fred Hutchinson Cancer Center

Understanding the association between features of different omics data types offers important imperative insights into biological mechanisms underlying human conditions and diseases, potentially leading to novel therapeutic opportunities. A common mode of analysis is to assess the association between pairs of features from different data types. However, this does not reflect the association between features whilst in the presence of all other features, leading to severe false positives. Adjustment is difficult due to high dimensionality and nonlinearity of relationships. To address this, we propose to calculate the scaled expected correlation (SECor) between pairs of features while marginalizing over all other features using nonparametric machine learning approaches. The SECor is asymptotically normal leading to easy p-value and interval construction. We demonstrate through simulations and applications to microbiome-metabolomics experiments that SECor well captures associations between pairs of features while in the presence of others.

11:20-11:45 AM

Integrative Analysis of Multiple Microbiome Datasets

Ni Zhao, Johns Hopkins University

Recent studies have highlighted the importance of human microbiota in our health and diseases. However, in many areas of research, individual microbiome studies often offer inconsistent results due to the limited sample sizes and the heterogeneity in study populations and experimental procedures. Integrative analysis of multiple microbiome datasets is necessary. However, statistical methods that incorporate multiple microbiome datasets and account for the study heterogeneity are not available in the literature. In this talk, I am going to discuss two recent developments from our lab that aim at integrative analysis of multiple microbiome datasets, one for the analysis of alpha diversity and one for the analysis of beta diversities. We applied these approaches to data from the HIV-reanalysis consortium, a collective effort that obtained all publicly available data on gut microbiome and HIV in December 2017, and obtained a coherent association of gut microbiome with HIV infection, and with MSM status (i.e. men who have sex with men).

11:45-12:10 PM

Quantifying Major Sources of Technical Variability in Microbiome Sequencing Lab Protocols

Ekaterina Smirnova, Virginia Commonwealth University

A major issue with horizontal harmonization of previously collected microbiome data is the large variation in the data processing through non-standardized methods. Microbiome study a complex process that starts with sample collection and storage, followed by transportation to a DNA extraction and sequencing lab, running bioinformatics pipeline to identify microbial taxa, and finally statistical analysis of a summarized taxonomic table. This leads to large differences in microbial taxa even for the replicate samples. We utilize the previously collected Microbiome Quality Control Project (MBQC) data that processed identical stool and artificial communities aliquots by 16 sample handling laboratories to quantify the major sources of technical variability on downstream statistical analysis. We re-process all sequences with the identical bioinformatics pipeline and rank the differences in alpha and beta diversity by sample handing protocols. The ultimate goal of this analysis is to inform the harmonization of previously collected microbiome studies and identify the major differences in microbiome sequencing protocols that have to be accounted for in the pulled studies analysis.

99. BAYESIAN METHODS IN DESIGN OF CLINICAL TRIALS

Organizer/Chair: Shirin Golchi, McGill University

10:30-10:55 AM

Designing a Bayesian Adaptive Clinical Trial Using Integrated Nested Laplace Approximations

Anna Heath, The Hospital for Sick Children; University of Toronto; University College London

Extensive simulations are required to design Bayesian adaptive trials, usually based on Markov Chain Monte Carlo simulations in non-conjugate settings. In practice, the computation cost of these simulations is a barrier to complex trial designs. We designed an adaptive perpetual trial that could stop for superiority or futility, declared based on posterior probabilities using the efficient Integrated Nested Laplace Approximations (INLA) algorithm to estimate the posterior in a proportional odds logistic regression. We calculated type I error and power across 64 scenarios that vary the posterior probability thresholds to stop for superiority and futility and the initial recruitment level before commencing adaptive analyses. Designs that maintained a type I error below 5%, a power above 80% and a feasible expected sample size, were evaluated across 19 different values for the odds ratios. Higher power was associated with larger initial sample sizes and expected sample size and higher thresholds for declaring futility. Two designs were selected for further evaluation. The efficiency INLA algorithm allowed us to optimise the trial design and evaluate the chosen design.

10:55-11:20 AM

Evaluation of Hybrid Controlled Trials that Leverage External Control Data and Randomization

Lorenzo Trippa, DFCI

Patient-level data from completed clinical studies or electronic health records can be used in the design and analysis of clinical trials. However, these external data can bias the evaluation of the experimental treatment when the statistical design does not appropriately account for potential confounders. In this work, we introduce a hybrid clinical trial design that combines the use of external control datasets and randomization to experimental and control arms, with the aim of producing efficient inference on the experimental treatment effects. Our analysis of the hybrid trial design includes scenarios where the distributions of measured and unmeasured prognostic patient characteristics differ across studies. Using simulations and datasets from clinical studies in extensive-stage small cell lung cancer and glioblastoma, we illustrate the potential advantages of hybrid trial designs compared to externally controlled trials and randomized trial designs.

11:20-11:45 AM

Elastic Priors to Dynamically Borrow Information from Historical Data in Clinical Trials

Ying Yuan, The University of Texas MD Anderson Cancer Center

Use of historical data and real-world evidence holds great potential to improve the efficiency of clinical trials. One major challenge is to effectively borrow information from historical data while maintaining a reasonable type I error and minimal bias. We propose the elastic prior approach to address this challenge. Unlike existing approaches, this approach proactively controls the behavior of information borrowing and type I errors by incorporating a well-known concept of clinically significant difference through an elastic function. The elastic prior approach has a desirable property of being information borrowing consistent, that is, asymptotically controls type I error at the nominal value, no matter that historical data are congruent or not to the trial data. Our simulation study that evaluates the finite sample characteristic confirms that, compared to existing methods, the elastic prior has better type I error control and yields competitive or higher power.

11:45-12:10 PM

Avoiding Bayes-Frequentist Clinical Trial Crosstalk

Janet Wittes, Wittes LLC

Physicians exposed to Bayesian thinking for the first time often find that it answers a great many of their concerns about the statistical paradigms they have been taught. No more dealing with the abstruse definition of a confidence interval. No more struggling with what a p-value means (or indeed whether to use p-values at all). No more worrying about the alpha-police hovering over every additional analysis. No more penalty for looking at data again and again. All this, and heaven too? Why, they ask, has no one ever told us about this? But there is a downside: when explained that the methods are very hard to describe because they require extensive simulation, they may ask, “But what does this get me that I couldn’t have gotten with the methods I know?” And “How do I know that the simulations address what needs to be addressed.” This talk deals with how we statisticians should introduce Bayes methodology in a way that helps the newcomer to Bayes understand the philosophical difference between Bayesian and frequentist thought and that shows parallels between the two types of inferences.

100. ADVANCED METHODS FOR ANALYZING LARGE-SCALE NEUROIMAGING DATA FROM NATIONWIDE CONSORTIUMS FOR MENTAL HEALTH RESEARCH

Organizer: Ying Liu, Columbia University
Chair: Younggeun Kim, Columbia University

10:30-10:55 AM

Statistical Challenges for Data Integration in Large Neuroimaging Cohort Studies

Haochang Shou, University of Pennsylvania

With the increasing need for big data analytics in medical imaging, integrating data from multiple study sites and various biological domains has become critical to better understanding complex human diseases. For example, large-scale observational studies often collected multiple modalities of measurements such as imaging, mobile health, and survey data from samples recruited over several clinical centers. Modeling and statistical inference of such data are particularly challenging due to the existence of site differences caused by unwanted technical variations that could mask the biological associations of interest, and those data modalities might be collected with different data types, dimensions and distributions. In this talk, we will discuss several most recent developments in statistical harmonization methods in large neuroimaging studies under various data modalities. We then introduce a novel distance-based regression model, which we refer to as Similarity-based Multimodal Regression (SiMMR), that enables simultaneous regression of multiple modalities through their distance profiles.

10:55-11:20 AM

Covariate Informed Identifiable Variational Autoencoder to learn Representation from Brain Imaging measures

Ying Liu, Columbia University

The recently proposed identifiable variational autoencoder (iVAE) framework provides a promising approach for learning latent independent components (ICs).Our motivating example is the Adolescent Brain Cognitive Development Study. When implementing the method to brain imaging measures from ABCD study, the existing i-VAE method encountered a degeneration problem. We show that iVAEs could have local minimum solution where observations and the approximated ICs are independent given covariates-- a phenomenon we referred to as the posterior collapse problem of iVAEs. To overcome this problem, we develop a new approach, covariate-informed iVAE (CI-iVAE) by considering a mixture of encoder and posterior distributions in the objective function. In doing so, the objective function prevents the posterior collapse, resulting latent representations that contain more information of the observations. Besides application to ABCD datasets, we present experiments on simulation datasets, EMNIST, Fashion-MNIST, and a large-scale brain imaging dataset demonstrate the effectiveness of our new method.

11:20-11:45 AM

Covariate-Modulated Bayesian Model for Whole-Brain Associations

Wesley Thompson, Laureate Institute for Brain Research

Neuroimaging studies have begun the process of expanding to much larger, population-valid sampling frames. Large, population-representative samples are beginning to provide the basis for a much-needed correction to the problems of small samples and can provide a more accurate picture of brain-behavior association sizes. Importantly, statistical significance by itself is not sufficient to characterize brain-behavior associations in the presence of widely distributed but very small effects. In this talk, we will discuss a novel whole-brain approach to assessing brain-behavior associations, using a Bayesian variance components model. The associations of voxels/vertices are allowed to differ by annotations, giving a "covariate-modulated" variance components analysis that can be used to assess the relative importance of, e.g., brain networks or patterns of gene expression across the brain. We demonstrate this new method by applying it to structural imaging data from the Adolescent Brain Cognitive Development (ABCD) Study, consisting of n=12,000 participants with longitudinal brain imaging.

11:45-12:10 PM

DISCUSSANT

Yuanjia Wang, Columbia University

101. NEW DEVELOPMENT OF APPROACHES IN THE FRONTIERS OF GENOMIC DATA SCIENCE

Organizer/Chair: Chad He, Fred Hutchinson Cancer Center

10:30-10:55 AM

Efficient SNP-Based Heritability Estimation Using Gaussian Predictive Process in Large-scale Cohort Studies

Saonli Basu, University of Minnesota

With the advent of high throughput genetic data, there have been attempts to estimate heritability from genome-wide SNP data on a cohort of distantly related individuals. Linear mixed models (LMMs) are used to estimate this heritability parameter. These LMM approaches assume the total variance to be comprised of genetic and environmental components. Heritability is the ratio of the genetic variance over the total variance of the trait. Fitting such an LMM in large-scale cohort studies and estimating these variance parameters, however, is tremendously challenging due to high dimensional linear algebraic operations. In our proposed work, we simplify the LMM by unifying the concept of Genetic Coalescence and Gaussian Predictive Process, thereby greatly alleviating the computational burden. Our proposed approach PredLMM has much better computational complexity than most of the existing packages and thus, provides an efficient alternative for estimating heritability in large-scale cohort studies. We illustrate our approach with extensive simulation studies to estimate the heritability of multiple quantitative traits from the UK Biobank cohort.

10:55-11:20 AM

Depth Normalization of Small RNA Sequencing: Using Data and Biology to Select a Suitable Method

Li-Xuan Qin, Memorial Sloan Kettering Cancer Center

Deep sequencing has become the most popular tool for transcriptome profiling in cancer research and biomarker studies. Similar to other high through-put profiling technologies such as microarrays, sequencing also suffers from systematic non-biological artifacts that arise from variations in experimental handling. A critical first step in sequencing data analysis is to “normalize” sequencing depth, so that the data can be comparable across the samples. A plethora of analytic methods for depth normalization has been proposed, and different normalization methods may lead to different analysis results with no method found to work systematically best. Currently, it is often up to the data analyst to choose a method based on personal preference and convenience. We developed a data-driven and biology-motivated approach to more objectively guide the selection of a depth normalization method for the data at hand. We assessed the performance of this approach using a unique pair of data sets for the same set of tumor samples that we previously collected. We then applied it to miRNA-sequencing data of 32 cancer types in the Cancer Genome Atlas.

11:20-11:45 AM

A Summary Statistics-based Method for Integrating Functional Information into Genetic Association Analysis of Multivariate Traits

Li Hsu, Fred Hutchinson Cancer Center

Genome-wide association studies (GWAS) have identified tens of thousands of genetic variants, but together they explain only a fraction of heritability, suggesting a much larger sample size is needed. However, it becomes increasingly difficult to achieve this. Recently, efforts have been devoted to improve power by leveraging the functional information of genetic variants and the information from multiple related phenotypes. In this talk, I will present a GWAS summary statistics-based method for testing genetic association of multivariate traits with a set of variants, incorporating the information on genetic regulation of molecular characteristics. We derived the multivariate trait test statistic, accounting for the correlation of multiple traits estimated from GWAS summary statistics. Simulation demonstrates that summary statistics-based p-values agree well with those from individual-level data, but with much faster computing speed. Importantly, a broad application of our method to GWAS is possible, as only summary statistics are required. We illustrate the method by assessing genetically predicted gene expression association with blood pressure and stroke.

11:45-12:10 PM

Gaussian Graphical Model-based Heterogeneity Analysis for Cancer Omics Data

Rong Li, Department of Biostatistics, Yale University

Heterogeneity is a hallmark of cancer omics studies. In "early" studies, mean, variance, and other marginal statistics have been commonly used for heterogeneity analysis. More recently, it has been found that network-based analysis can be more informative and may generate heterogeneity structures missed by previous analyses. Building on a series of recent GGM studies, we have further developed heterogeneity analysis methods that can accommodate additional effects, for example, from unknown latent variables or a set of high-dimensional confounders. In cancer omics studies, such methods can effectively remove the effects of known and unknown demographic and clinical factors as well as regulators (for example, methylation for gene expressions).

102. NEW METHODS IN THE CONTEXT OF CHEMICAL MIXTURES ON HEALTH OUTCOMES

Organizer: Rajeshwari Sundaram, National Institutes of Health
Chair: Abhisek Saha, National Institutes of Health

10:30-10:55 AM

Using Metrics of a Mixture Effect and Nutrition for Consideration in Causal Inference

Chris Gennings, Icahn School of Medicine at Mount Sinai

Environmental exposures (e.g., from consumer products, building materials, pesticides, water, air pollution) have been associated with adverse health effects in humans. Single chemical experimental studies demonstrate causal links between chemicals and outcomes, but such studies do not represent human exposure. Our objective is to demonstrate the use of an index as a metric towards causally linking human exposure to health outcomes. We use both an empirically-weighted exposure index to summarize the mixture effect of the joint action of chemical mixtures and a nutrition index to address both adverse exposures and beneficial dietary nutrients. We found that in a pregnancy cohort early prenatal exposure to a mixture of EDCs and poor nutrition are associated with a decrease in cognitive functioning at age 7 years. Reducing the EDC exposure or improving dietary nutrition using these metrics, the counterfactuals in a step towards causal inference using g-computation methods, indicate significant improvement in cognitive function. Evaluation of such a strategy may support decision makers for risk management of EDCs and individual choices for improving dietary nutrition.

10:55-11:20 AM

A Comparison of Statistical Methods for Estimating Interactions Among Chemical Mixtures

Paul Albert, National Cancer Institute

Estimating the interactions between chemical mixture components is often of interest in epidemiologic studies. Motivated by a study examining the effects of chemical exposures on lung cancer in China, we compared different approaches for incorporating interactions in a case-control study. Specifically, we compared Bayesian Kernal machine regression (BKMR) and Bayesian LASSO to two recently developed approaches. The first is a latent functions approach where main and interactions are estimated assuming two separate sets of unobserved functions. The second is a Bayesian shrinkage approach that incorporates the hierarchical principle which assumes it is unjlikely that there are interactions without the presence of main effects. We show that the comparison of these different methods on the lung cancer data example provides interesting insight into biological mechanism of the disease process.

11:20-11:45 AM

Infinite Hidden Markov Models for Multiple Multivariate Time Series with Missing Data

Ander Wilson, Colorado State University

Exposure to air pollution is associated with increased morbidity and mortality. Recent technological advancements permit the collection of time-resolved personal exposure data on air pollution mixtures. Such data are often incomplete with missing observations and exposures below the limit of detection, which limit their use in health effects studies. We develop an infinite hidden Markov model for multiple asynchronous multivariate time series with missing data. Our model is designed to include covariates that can inform transitions among hidden states. We implement a Bayesian multiple imputation algorithm to impute missing data both missing at random and below a limit of detection. We validate our model in both a simulation study and in a case study using the Fort Collins Commuter Study, which estimated ten-second resolution personal exposure data to air pollution mixtures. In a case study, we describe the inferential gains obtained from our model including improved imputation of missing data and the ability to identify shared patterns in activity and exposure among repeated sampling days for individuals and among distinct individuals.

11:45-12:10 PM

Simultaneous Variable Selection and Inference in Joint Model with Skewed Longitudinal Process and Discrete Time-to-Event with Application to Assessing Environmental Mixtures with Fecundity

Rajeshwari Sundaram, Eunice Kennedy Shriver National Institutes of Health, NIH

Assessing multi-pollutant association with health outcomes are of considerable interest. Much effort has been focused on assessing it in the context of single outcome, but limited literature exists in the context of joint modeling. We focus on addressing this issue in the context of joint modeling of menstrual cycle lengths (MCL) and time-to-pregnancy (TTP). Previous literature has proposed a skewed distribution to model the MCL and a discrete survival model for the TTP. Corresponding to this, we consider joint modeling of a skewed longitudinal process and a discrete survival time model with a shared frailty. We consider the approach of variable selection using elastic-net that allows one to identify important drivers of the mixtures of pollutants in the context of joint modeling. For the problem, a class of penalized likelihood-based procedures will be developed for simultaneous variable selection and estimation of relevant covariate effects for both longitudinal and survival time variables of interest. The proposed approach is evaluated extensively through simulations and an analysis of a prospective pregnancy LIFE Study.

103. CONTRIBUTED PAPERS: STATISTICAL GENETICS

Chair: Shijia Bian, Department of Biostatistics and Bioinformatics, Emory University

10:30-10:45 AM

Sparse Models for Structured Matrix-Valued Data from High-Throughput Studies

Saunak Sen, University of Tennessee Health Science Center

High-throughput biological data are often in the form of a large matrix with covariates annotating both its rows and columns. Matrix linear models provide a convenient way for modeling such data. In many situations, sparse estimation of these models is desired. We present fast, general methods for fitting sparse matrix linear models to structured high-throughput data. We induce model sparsity using an L1 (and elastic net) penalty and consider the case when the response matrix and the covariate matrices are large. Due to data size, standard methods for estimation of these penalized regression models fail if the problem is converted to the corresponding univariate regression scenario. By leveraging matrix properties in the structure of our model, we developed several fast estimation algorithms each with their strengths and weaknesses. We evaluate our method?s performance on simulated data, E. coli chemical genetic screening data and two Arabidopsis genetic datasets with multivariate responses. Our algorithms have been implemented in the Julia programming language and are available at https://github.com/senresearch/MatrixLMnet.

10:45-11:00 AM

Deconvolution Analysis of Cell Type Expression from Bulk Tissues by Integrating with Single-cell Expression Reference

Yutong Luo, Department of Biostatistics, Bioinformatics, and Biomathematics, Georgetown University Medical Center

To understand phenotypic variations and key factors which affect disease susceptibility of complex traits, it is important to decipher cell type tissue compositions. To study cellular compositions of bulk tissue samples, one can evaluate cellular abundances and cell type specific gene expression patterns from the tissue transcriptome profiles. We develop both fixed and mixed models to reconstruct cellular expression fractions for bulk-profiled samples by using reference single-cell (sc) RNA-sequencing (RNA-seq) reference data. In benchmark evaluations, the mixed effect models provide similar results as CIBERSORTx, which is a well-known and reliable procedure to reconstruct cell-type-specific gene expression profiles. In real data analysis, the mixed effect models outperform or perform similarly as CIBERSORTx. The mixed models perform better than the fixed models in both benchmark evaluations and data analysis. In simulation studies, we show that if the heterogeneity exists in a scRNA-seq data, it is better to use mixed models with heterogeneous mean and variance-covariance. The proposed mixed models provide a complementary tools to dissect bulk tissues using scRNA-seq data.

11:00-11:15 AM

Canopy2: Tumor Phylogeny Inference Using Bulk DNA and Single-Cell RNA Sequencing

Ann Marie Weideman, Department of Biostatistics, University of North Carolina at Chapel Hill

Single-cell sequencing is a powerful platform to assess tumor heterogeneity and track cancer evolution. We propose Canopy2, a statistical and computational method for tumor phylogeny inference using single-nucleotide variants derived from bulk DNA and single-cell RNA sequencing. Canopy2 samples from a joint probability distribution involving a mixture of binomial and beta-binomials, specifically chosen to account for the sparsity and stochasticity of the single-cell data. Compared to existing methods, Canopy2 uses both bulk and single-cell data (versus one or the other) to update the joint posterior. Additionally, Canopy2 demystifies the sources of zeros in the single-cell data and separates non-cancerous (cells without mutations), stochastic (mutations not expressed due to bursting), and technical (expressed mutations not picked up by sequencing) zeros. Canopy2's performance was assessed by simulations, and the methodology was applied to bulk and single-cell breast cancer and glioblastoma datasets, with single-cell DNA sequencing data for validation. We show that Canopy2 successfully infers tumor phylogenetic history and conducts mutational profiling of tumor subpopulations.

11:15-11:30 AM

Genetic Association Analysis of Multiple Binary and Quantitative Traits Considered Jointly

Yi Wei, University of Chicago

In genetic association analysis, when data are available on multiple related traits or measurements, power can potentially be gained by analyzing them jointly. However, when some of the traits considered are binary, use of methods designed for quantitative traits could result in power loss when covariate effects are important. We propose a new method for multi-trait mapping of a combination of binary and quantitative phenotypes, which is appropriate when the number of traits to be jointly analyzed is not large. The method is based on a mixed-effects quasi-likelihood framework, which captures the dichotomous nature of the binary traits and can incorporate covariates and population structure. We test for association based on a retrospective approach, which is robust to misspecification of the phenotype model. When binary traits are included in the analysis, parameter estimation presents additional challenges beyond those for the quantitative trait case. In estimating the correlation matrices, we use a recently proposed parametrization that can be viewed as a multivariate generalization of Fisher's Z-transformation of a single correlation.

11:30-11:45 AM

Hardy-Weinberg Equilibrium Test Accounting for Population Structure and Genetic Relatedness

Derek Shyr, Harvard T.H. Chan School of Public Health

Large-scale cohort studies such as the Center for Common Disease Genomics have integrated deep whole-genome sequencing and other omics data with clinical data. Assessing the Hardy-Weinberg Equilibrium (HWE) assumption appropriately for these datasets is integral to quality control procedures; however, this is challenging due to ancestral heterogeneity and genetic relatedness of the samples. Currently, there are no existing methods that incorporate both population structure and genetic relatedness when testing for HWE. We propose a novel HWE test using the generalized estimating equation that accounts for population structure with principal components and the relationship among samples with a family-specific genotype correlation matrix. Our results demonstrate that ignoring population structure and relatedness when evaluating HWE inflates the false-positive rates drastically. Compared to other methods, our approach controls for type-I error the best while maintaining high power. Our implementation is scalable and practical such that HWE tests can be performed efficiently across millions of markers and over a hundred thousand samples.

11:45-12:00 PM

Improving Polygenic Risk Prediction in Admixed Populations by Explicitly Modeling Ancestral-Specific Effects via GAUDI

Quan Sun, University of North Carolina at Chapel Hill

Polygenic risk scores (PRS) have shown successes in clinics, but their performance in non-European individuals remains unsatisfying. Multiple methods have been developed to improve PRS in diverse ancestral groups, but they focus mostly on participants with primarily one ancestry, not considering admixed individuals with distinct ancestries for different segments of their genomes. Here, we propose GAUDI, a novel penalized-regression-based method specifically designed for admixed individuals. By adopting a modified lasso framework and balancing between sparsity and fusion, GAUDI explicitly models ancestry-specific effects and jointly estimates ancestry-shared effects, borrowing information across segments with shared ancestry in admixed genomes. We demonstrate the advantage of GAUDI through comprehensive simulations. Leveraging data from the Women's Health Initiative study, we show GAUDI improves PRS prediction in African Americans by 64% - 760% relatively compared to other methods. We believe GAUDI will be a valuable tool to mitigate disparities in PRS performance across ancestral groups.

12:00-12:15 PM

Integrative Cross-Omics and Cross-Context Analysis Elucidates Molecular Links Underlying Genetic Effects on Complex Traits

Yihao Lu, The University of Chicago

The proliferation of association results from large-scale genetic and genomic studies provides opportunities and challenges to elucidate the disease/trait mechanisms and their operating cellular contexts. Motivated by a multi-tissue multi-omics analysis using genetics, methylome and transcriptome data from the Genotype-Tissue Expression (GTEx) project, we propose a method, X-ING (Cross-INtegrative Genomics), for cross-omics and cross-context integrative analysis of summary-level data. X-ING takes as input the statistic matrices from multiple omics studies, each with multivariate contexts. It models the latent binary association status of each statistic and captures the omics-shared and context-shared major patterns in a hierarchical Bayesian model. Our analysis of cis-genetic effects on methylome and transcriptome from GTEx characterizes the tissue/omics-effect sharing patterns. The analysis of trans-genetic effects demonstrates enrichment of trans-associations in many disease/trait-relevant tissues. Many associations identified by X-ING are replicated in external data, with higher replication rates for multi-tissue or multi-omics effects.

104. CONTRIBUTED PAPERS: DECISION MAKING, REFERENCE/TOTAL EFFECT ESTIMATION IN META-ANALYSIS

Chair: Fan Bu, University of California, Los Angeles

10:30-10:45 AM

A Bayesian Nonparametric Meta-Analysis Model for Estimating the Reference Interval

Wenhao Cao, Division of Biostatistics, University of Minnesota

A reference interval refers to the normative range for measurements from a healthy population. It plays an important role in laboratory testing, as well as differentiating the healthy from diseased patients. The reference interval based on a single study has serious limitations and the results might not be appropriate for a broader population. Meta-analysis can provide a general reference interval based on the overall population by combining results from multiple studies. However, the normal distribution of underlying study-specific means, and equal within-study variances assumption, which are extensively used in existing methods, are too strong and could be violated. We use a Bayesian nonparametric model with more flexible assumptions to extend the application of random effects meta-analysis for estimating reference intervals. We illustrate through simulation and real data analysis the robustness of our proposed approach under the violation of normal assumptions and equal-within-study variances assumption.

10:45-11:00 AM

RIMeta: An R Shiny Tool for Estimating the Reference Interval from a Meta-Analysis

Ziren Jiang, University of Minnesota

A reference interval, or an interval in which a pre-specified proportion of measurements from a healthy population are expected to fall, is used to determine whether a person's measurement is typical of a healthy individual. For a specific biomarker, multiple published studies may provide data collected from healthy participants. A reference interval estimated by combining the data across these studies is typically more generalizable than a reference interval based on a single study. Methods for estimating reference intervals from random effects meta-analysis and fixed effects meta-analysis have been recently proposed and implemented using R software. We present an R Shiny tool, RIMeta, implementing these methods, which allows users not proficient in R to estimate a reference interval from a meta-analysis using aggregate data (mean, standard deviation, and sample size) from each study. RIMeta provides users a convenient way to estimate a reference interval from a meta-analysis and to generate the reference interval plot to visualize the results.

11:00-11:15 AM

Bayesian Hierarchical Models for Multivariate Meta-Analysis of Diagnostic Tests in the Absence of a Gold Standard with an Application to SARS-CoV-2 Infection Diagnosis

Zheng Wang, Division of Biostatistics, School of Public Health, University of Minnesota

When evaluating a diagnostic test, it is common that a gold standard is absent. One example is the diagnosis of SARS-CoV-2 infection using saliva sampling or nasopharyngeal swabs. Without a gold standard, a pragmatic approach is to postulate a reference standard, defined as positive if either test is positive, or negative if both are negative. However, this pragmatic approach may overestimate sensitivities because subjects infected with SARS-CoV-2 may still have double-negative test results even when both tests exhibit perfect specificity. To address this limitation, we propose a Bayesian hierarchical model for simultaneously estimating the sensitivities, specificities, and disease prevalence in the absence of a gold standard. The proposed model allows adjusting for study-level covariates. We evaluate the model performance using a worked example based on a recently published meta-analysis on the diagnosis of SARS-CoV-2 infection and extensive simulations. Compared with the pragmatic reference standard approach, we demonstrate that the proposed Bayesian method provides a more accurate evaluation of prevalence, specificities, and sensitivities in a meta-analytic framework.

11:15-11:30 AM

Improving Estimation of Total Effects in Meta-Analysis

Colleen Chan, Yale University

Meta-analyses summarize evidence about an association across multiple sources of information, increasing statistical power and exploring sources of heterogeneity. Yet, meta-analyses may neglect the complex causal structure behind an association, failing to distinguish between total and direct effects in the presence of a mediator. When the total effect is unavailable, some meta-analyses include the direct effect in place of the total effect, biasing the summary of the association towards the null. We develop methods to estimate point and interval estimates of the mediation proportion and total effect in this setting, filling an important methodological gap in existing evaluation approaches. In addition to reducing bias, by leveraging a summary mediation proportion whose estimator is developed here, our method is able to include a wider range of studies in the meta-analysis, thus providing more efficient estimates under certain conditions. The methodology is illustrated by a meta-analysis of sugar-sweetened beverage consumption in relation to type 2 diabetes incidence. We also present an R package that implements our proposed methods.

11:30-11:45 AM

Non-Greedy Tree-Based Learning for Estimating Global Optimal Dynamic Treatment Decision Rules with Continuous Treatment Dosage

Chang Wang, University of Michigan

Dynamic treatment regime (DTR) plays a critical role in precision medicine when assigning patient-specific treatments at multiple stages and optimizing a long-term clinical outcome. However, most of existing work about DTRs have been focused on categorical treatment scenarios, instead of continuous treatment options. Also, the performances of black-box algorithm and regular tree learning methods are lack of interpretability and global optimality respectively. We propose a non-greedy global optimization method for dose search, namely Global Optimal Dosage Tree-based learning (GoDoTree), which combines a robust estimation of the counterfactual outcome with an interpretable and non-greedy decision tree for estimating the global optimal dynamic dosage treatment regime in a multiple-stage setting. GoDoTree recursively estimates how the counterfactual outcome mean depends on a continuous treatment dosage using doubly robust estimators at each stage, and optimizes the stage-specific decision tree in a non-greedy way. We conduct simulation studies to evaluate the finite sample performance of proposed method and apply it to a real data application for optimal warfarin dose finding.

11:45-12:00 PM

RISE: Robust Individualized Decision Learning with Sensitive Variables

Xiaoqing Tan*, University of Pittsburgh

This paper introduces RISE, a ?robust individualized decision learning framework with sensitive variables, where sensitive variables are collectible data and important to the intervention decision, but their inclusion in decision making is prohibited due to reasons such as delayed availability or fairness concerns. A naive baseline is to ignore these sensitive variables in learning decision rules, leading to significant uncertainty and bias. To address this, we propose a decision learning framework to incorporate sensitive variables during offline training but not include them in the input of the learned decision rule during model deployment. Specifically, from a causal perspective, the proposed framework intends to improve the worst-case outcomes of individuals caused by sensitive variables that are unavailable at the time of decision. Unlike most existing literature that uses mean-optimal objectives, we propose a robust learning framework by finding a newly defined quantile- or infimum-optimal decision rule. The reliable performance of the proposed method is demonstrated through synthetic experiments and three real-data applications.

12:00-12:15 PM

Kullback-Leibler-Based Discrete Failure Time Models for Integration of Published Prediction Models with New Time-To-Event Dataset

Di Wang*, University of Michigan

Prediction of time-to-event data often suffers from rare event rates, small sample sizes, high dimensionality and low signal-to-noise ratios. Incorporating published prediction models from large-scale studies is expected to improve the performance of prognosis prediction on internal individual-level time-to-event data. However, existing integration approaches typically assume that underlying distributions from the external and internal data sources are similar, which is often invalid. To account for challenges including heterogeneity, data sharing, and privacy constraints, we propose a discrete failure time modeling procedure, which utilizes a discrete hazard-based Kullback-Leibler discriminatory information measuring the discrepancy between the published models and the internal dataset. Simulations show the advantage of the proposed method compared with those solely based on the internal data or published models. We apply the proposed method to improve prediction performance on a kidney transplant dataset from a local hospital by integrating this small-scale dataset with published survival models obtained from the national transplant registry.

105. CONTRIBUTED PAPERS: BIOPHARMACEUTICAL RESEARCH METHODS

Chair: Thomas Lumley, University of Auckland

10:30-10:45 AM

A New Family of Covariate-Adjusted Response-Adaptive Randomization Procedures for Precision Medicine

Jiaqian Yu, The George Washington University

In most clinical trials, patients accrue sequentially and need to be assigned to different treatment groups. With the development of precision medicine, information about biomarkers (covariates) is usually available and should be included in the randomization procedure. In literature, covariates (biomarkers) are classified into predictive and prognostic covariates according to their roles. Under this setting, we propose a new family of adaptive randomization procedures that can not only assign more patients to better treatments according to the predictive covariates, but also balance prognostic covariates. Theoretical properties of the proposed procedures are derived. The advantages of the proposed procedures are demonstrated by numerical studies as well as real examples.

10:45-11:00 AM

Analysis of Prostate Histology Image for Cancer Detection and Grading

Mohammad Samsul Alam, North Carolina State University

Histology images are the cornerstone for confirming and understanding cancer. The usual practice is to use Hematoxylin and Eosin (H&E) staining for highlighting different tissue components in this type of imaging. These stains highlight epithelial cells with dark purples, whereas the glandular areas with white. When cancer starts growing and spreading, epithelial cells mutate irregularly and break their regular pattern, congregating closely on the cytoplasm around the glandular areas. This behavior changes the spatial distribution of epithelial and glandular mechanisms. By exploiting this change, we develop a statistical approach for predicting and grading prostate cancer. Our approach uses spatially indexed vector-valued function data methods to summarize the image patches. We use a secondary dimension reduction to extract the main features in a way that captures the spatial information existing in the patches. These resulting summaries are then used to identify the cancerous images and grade the cancer severity of the patches. We present the results on PESO data set that contains H&E stained patches collected from whole slide images of prostate tissue.

11:00-11:15 AM

Heterogenous Treatment Effect of Patients with Acute Myeloid Leukemia

Gege Gui, Johns Hopkins University

Randomized clinical trials for acute myeloid leukemia have shown that two groups of pre-transplant conditioning regimens (MAC vs RIC/NMA) influenced patient clinical outcomes. However, the trial results estimate the average treatment effect in a specific patient population that may or may not be generalizable to a broader group of clinics where most patients receive their care. We obtained clinical and genomic data of 1075 adult patients from 111 sites of Center for International Blood and Marrow Transplant Research from 2013-2019 to study the effect of pre-transplant conditioning regimes on relapse rate and time to relapse in a general patient population. These data comprise an observational study because treatment assignment was not randomized. This paper presents a case study comparing regression and machine learning methods to: (1) adjust for non-random treatment assignment; and (2) estimate conditional average treatment effects as a function of baseline demographic and genomic markers. We demonstrate the value of simulation to tailor statistical methods to the particular scientific study data.

11:15-11:30 AM

Analyzing Randomized Experiments Subject to Outcome Misclassification via Integer Programming

Siyu Heng, Department of Biostatistics, School of Global Public Health, New York University

Results from randomized experiments (trials) can be severely distorted by outcome misclassification, such as from measurement error or reporting bias in binary outcomes. All existing approaches to outcome misclassification rely on some data-generating (super-population) model and therefore may not be applicable to randomized experiments without additional assumptions. We propose a model-free and finite-population-exact framework for randomized experiments subject to outcome misclassification. A central quantity in our framework is "warning accuracy," defined as the threshold such that the causal conclusion drawn from the measured outcomes may differ from that based on the true outcomes if the outcome measurement accuracy did not surpass that threshold. We show how learning the warning accuracy and related concepts can benefit a randomized experiment subject to outcome misclassification. We show that the warning accuracy can be computed efficiently (even for large datasets) by adaptively reformulating an integer program with respect to the randomization design. We apply our framework to a large randomized clinical trial for the prevention of prostate cancer.

11:30-11:45 AM

A Benchmark Effective Sample Size to Measure Information Borrowing in Hybrid Designs

Evan Kwiatkowski, MD Anderson Cancer Center

Hybrid designs are an important approach to leveraging historical control data to reduce the sample size of standard randomized controlled trial (RCT) designs by utilizing hybrid controls, which consist of concurrent controls and augmented controls “borrowed” from historical data. We develop a pragmatic method to determine the effective sample size of the hybrid controls by benchmarking the RCT design that the hybrid design targets, referred to as the benchmark effective sample size (BESS). The BESS of the hybrid controls is determined by the number of RCT controls that would be needed to match the conditional power curve from the hybrid design as compared to the power achieved from the benchmark RCT design. The BESS has several desirable properties. First, it depends on the observed data (e.g., observed effect size), rather than an average over hypothetical datasets. Consequently, BESS inherently adjusts for prior-data conflict. Second, by benchmarking to a RCT design, BESS does not rely on the notion of a chosen non-informative/vague prior to quantify the effective sample size. Third, BESS can be computed with any selected prior.

11:45-12:00 PM

A Comparison of Statistical Tests for Prioritized Time-to-Event Outcomes

Jingyi Lin, Boston University

Many clinical trials evaluate treatment effects based on a composite time-to-event outcome. A single composite measure only captures a limited aspect of the entire disease burden. This paper compares the performance of both novel and traditional statistical methods for analyzing multiple prioritized time-to-event outcomes. We categorized existing methods into three groups: (1) methods that respect a hierarchy of clinical importance, e.g. the generalized pairwise comparison and the restricted mean time in favor of treatment; (2) methods that combine component-wise treatment effects, e.g. the O'Brien test and the Wei-Lachin test; (3) methods that do not differentiate among components and lead to an indecomposable composite measure, e.g. the max-combo test and the area under the cumulative event count curve. Our simulation design features heterogeneous component-wise treatment effects. The simulation also involves non-proportional hazard settings with early to late treatment effects. We compared the empirical power and the corresponding effect size of each test under the dynamically-designed simulation study to understand the pros and cons of different types of tests.

106. CONTRIBUTED PAPERS: EPIDEMIOLOGICAL STUDIES AND CAUSAL INFERENCE

Chair: Gary Hettinger, University of Pennsylvania

10:30-10:45 AM

Bayesian Modeling of Synergistic and Antagonistic Interactions in Assessing Health Effects of Mixtures of Exposures

Shounak Chattopadhyay, Duke University

There is abundant interest in assessing the joint effects of multiple exposures on human health. This is often referred to as the mixtures problem in environmental epidemiology and toxicology. Classically, studies have examined the adverse health effects of different chemicals one at a time, but there is concern that certain chemicals may act together to amplify each other's effects. Such amplification is referred to as synergistic interaction, while chemicals that inhibit each other's effects have antagonistic interactions. Current approaches for assessing the health effects of chemical mixtures do not explicitly consider synergy or antagonism in the modeling, instead focusing on either parametric or unconstrained nonparametric dose response surface modeling. We propose Synergistic Antagonistic Interaction Detection (SAID), a Bayesian approach which identifies pairwise synergistic or antagonistic interactions, while providing variable selection decisions for each component. This framework is evaluated relative to existing approaches using simulation experiments and an application to data from NHANES.

10:45-11:00 AM

Approaches to Estimate Bidirectional Causal Effects Using Mendelian Randomization

Jinhao Zou, Department of Biostatistics, The University of Texas MD Anderson Cancer Center

Mendelian Randomization (MR) is a framework of using genetic variants as instrumental variables (IVs) to examine the causal effect of exposure on outcome in observational data. Statistical methods based on unidirectional MR (UMR) are widely used in current studies. When bidirectional causality between two phenotypes is investigated, the UMR methods are applied bidirectionally to estimate causal effects in each direction. However, bidirectional causality between two phenotypes leads to a feedback loop between them, which biases the estimations using UMR methods. We proposed two novel methods for the bidirectional MR (BMR) model with a feedback loop: BiRatio and BiLIML. We evaluated the new BMR methods through simulations with both unidirectional and bidirectional causality scenarios. Our simulations showed that BiRatio and BiLIML methods provide accurate estimations using strong IVs. Using weak IVs, the BiLIML method provides the most accurate estimations. Applying the proposed methods to data from the Multi-Ethnic Study of Atherosclerosis, our results revealed the bidirectional causal relationship between obesity and diabetes.

11:00-11:15 AM

Tailoring Capture-Recapture Methods to Estimate Registry-Based Case Counts Based on Error-Prone Diagnostic Signals

Lin Ge, Emory University, Department of Biostatistics and Bioinformatics

Surveillance research is important for epidemiological monitoring of disease prevalence. Motivated from ongoing efforts to identify recurrent cases based on the Georgia Cancer Registry, we extend recently proposed ?anchor stream? sampling design and estimation methodology. Our approach offers a more efficient and defensible way to traditional CRC methods by leveraging a relatively small random sample of participants. The key extension developed here accounts for the common problem of false positive or negative diagnostic signals from one or more of the existing data streams. In particular, we show that the design only requires documentation of positive signals in the surveillance streams, and permits valid estimation of the true case count based on an estimable positive predictive value. We borrow ideas from the multiple imputation paradigm to provide accompanying standard errors, and develop a Bayesian credible interval approach that yields favorable frequentist coverage properties. We demonstrate the proposed methods through simulation studies, and provide a data example about the breast cancer recurrences from the CRISP database.

11:15-11:30 AM

Efficient Study Designs and Analysis Methods for Longitudinal Binary Data: An Application to the Lung Health Study

Chiara Di Gravio*, Vanderbilt University

In modern epidemiological studies, researchers might be interested in understanding the relationship between a readily available longitudinal binary outcome and a novel exposure. When exposure ascertainment costs limit the sample size, two-phase studies are a pragmatic solution that allows researchers to target informative individuals for exposure ascertainment and increase the precision associated with estimating time-varying and/or time-fixed exposure associations. In this talk, we introduce a novel class of residual-based two-phase designs that selects informative individuals using all information available on outcome and covariates. Additionally, we propose a semi-parametric analysis approach that efficiently uses all available data and estimates parameters using a numerically stable and computationally efficient EM algorithm. We examine the finite sample operating characteristics of our approach through extensive simulation studies. We illustrate the usefulness of the proposed designs and method in practice through an application that studies associations between a genetic marker and poor lung function using data from the Lung Health Study.

11:30-11:45 AM

Inverse Probability of Censoring Weighted Super Learner for Survival Prediction in Case-Cohort Studies

Haolin Li, Department of Biostatistics, University of North Carolina at Chapel Hill

In modern epidemiological studies of rare diseases, the case-cohort study design is widely used to reduce the cost and achieve the same efficiency as a cohort study. Previous works have focused on analyzing data from the case-cohort design based on a particular statistical model but few have discussed the survival prediction problem under such type of design. In this article, we propose an inverse probability of censoring weighted super learner algorithm for survival prediction in case-cohort studies. The algorithm has also been extended to generalized case-cohort studies. The proposed super learner algorithm is shown to have asymptotic model selection consistency and uniform consistency and demonstrates satisfactory finite sample performances. We also show that the super learners trained by data from case-cohort studies have better prediction accuracy than the ones trained by data from simple random sampling given the same sample sizes. Finally, we illustrate the use of our method in a case-cohort study conducted as part of the Atherosclerosis Risk in Communities Study.

11:45-12:00 PM

Estimating Time-Varying Direct Causal Excursion Effects with Longitudinal Binary Outcomes

Jieru Shi, University of Michigan, Ann Arbor

Construction of just-in-time adaptive interventions, such as prompts delivered by mobile apps to promote and maintain behavioral change, requires knowledge about the time-varying moderated effects to inform occasions of high or low treatment effects. Micro-randomized trials (MRT) have emerged as sequentially randomized designs to gather the requisite data for effect estimation. The existing literature (Qian et al., 2020; Boruvka et al., 2018; Dempsey et al., 2020) has defined a general class of causal estimands, referred to as causal excursion effects, to assess the time-varying moderated effects. However, the statistical literature remains scant on how to address potential between-cluster treatment effect heterogeneity and within-cluster interference in a sequential treatment setting. In this paper, based on a cluster conceptualization of the potential outcomes, we define a larger class of direct causal excursion effects for proximal and lagged binary outcomes, and propose a new inferential procedure under effect heterogeneity and interference. We provide theoretical guarantees of consistency and asymptotic normality of the estimator.

12:00-12:15 PM

Statistical Methods for Assessing Treatment Effects on Ordinal Outcomes Using Observational Data

Huirong Hu, University of Louisville

Average treatment effect (ATE) is used to measure the outcome difference if all patients would have been treated under one treatment versus all patients would have been treated under another treatment. Many statistical methods have been developed to estimate ATE when the outcome is continuous or binary. In many cases, ordinal outcome is often used. We may lose information if we consider ordinal outcomes as continuous variable. In this project, we propose model the ordinal outcome directly using the marginal structure model (MSM), and we estimate the average treatment effect by the superiority score of the outcome under one treatment group versus another. To adjust the confounding factors between treatment and outcome in an observational study, we apply the inverse probability of treatment weighting (IPTW) to obtain a weighted sample, where the covariates become balanced among different treatment groups. We then assess ATE based on the weighted sample. Extensive simulation studies are carried out to examine the performance of the proposed method. We also applied the proposed method to access the ATE from alcohol use disorders using Kentucky Medicaid data.

107. CONTRIBUTED PAPERS: EXPERIMENTAL DESIGN, CLASSIFICATION, PREDICTION

Chair: Garrett Frady, University of Connecticut

10:30-10:45 AM

Virtual Baseline Generators - Concept, Assumptions and Example

Nicolas Savy, Toulouse Institute of Mathematics

With the development of artificial intelligence approaches, the volume of databases is becoming a major issue in medical research. Unfortunately, in this context, databases are often modest in size. One answer to this problem could come from data augmentation techniques. This method is broken down into two steps: first a step of learning of the generative process of the data in a non-parametric framework or of calibrating a generative model of the data in the parametric case, then a step of simulation according to the learned (estimated) model. The objective of this generator called virtual baseline generator is to obtain profiles of patients similar to those of the learning base or on the contrary to obtain patients according to a particular but realistic profile. In this presentation we will walk through several database augmentation techniques. We will try to explain the underlying assumptions of these techniques in order to propose a relevant methodology for use. Finally, we will illustrate the performance on an example.

10:45-11:00 AM

Testing of Treatment Effect in Crossover Design with Intensive Measurements

Jianping Sun, UNC Greensboro

A crossover study is a longitudinal study in which subjects receive a sequence of different treatments during the different time periods. In a crossover design, the treatments are compared within subjects, and hence no subject effects exist in the comparison. However, the existence of potential carryover effect could make the crossover design biased and invalid. Traditionally, there is usually only one measure of outcome after each treatment period in a crossover design. With the new technology, such as wearable devices, the outcome variables could be measured continuously and intensively in many trials, which brings fresh feature and challenges to the traditional crossover design. Thus, in this talk, we demonstrate a novel angle of understanding crossover studies, especially the classic AB/BA design, by using intensive repeated measurements. We propose that repeated measures could lead to a brand-new way of dealing with carryover effects and other effects confounded with treatment through a longitudinal model with non-linear trend of treatment effect. Simulation studies and real data analysis are conducted to examine the performance of our proposed method.

11:00-11:15 AM

Optimal Sampling for Positive Only Electronic Health Record Data

Seong-ho Lee, The Pennsylvania State University

Identifying a patient?s disease/health status from electronic medical records is a frequently encountered task in EHR related research, and estimation of a classification model often requires a benchmark training data with patients? known phenotype statuses. However, assessing a patient?s phenotype is costly and labor intensive, hence a proper selection of EHR records as a training set is desired. We propose a procedure to tailor the best training subsample with limited sample size for a classification model, minimizing its mean squared phenotyping/classification error (MSE). Our approach incorporates ?positive only? information, an approximation of the true disease status without false alarm, when it is available. In addition, our sampling procedure is applicable for training a chosen classification model which can be misspecified. We provide theoretical justification on its optimality in terms of MSE. The performance gain from our method is illustrated through simulation and a real data example, and is found often satisfactory under criteria beyond mean squared error.

11:15-11:30 AM

Optimizing Disease Prevalence Estimation Using Adaptive Group Testing

Shamim Sarker, Radford University

Surveillance of infectious diseases can be efficiently performed by using group (pooled) testing. The benefits gained from group testing depend on the pool size used. Statistical methods have been developed in the literature to find optimal pool sizes. Unfortunately, those methods use either simpler pooling protocols or require the use of perfect diagnostic assays. The goal of our work is to overcome these limitations and to provide a general optimization technique. We study the estimation efficiency and cost efficiency of a disease prevalence estimator calculated from group testing data. Then the optimal pool size is determined by minimizing the efficiency measures. To mitigate the dependence of optimization on an a priori estimate of the true disease prevalence, we take a multistage adaptive pooling approach. We find that a substantial gain in estimator efficiency can be realized using our work even when the a priori estimate is misspecified. A software application using the shiny package in R is also provided for ease of implementation of our work.

11:30-11:45 AM

Hierarchical Neyman-Pearson Classification for Prioritizing Severe Disease Categories in COVID-19 Patient Data

Lijia Wang, University of Southern California

COVID-19 has a spectrum of disease severity, ranging from asymptomatic to requiring hospitalization. Providing appropriate medical care to severe patients is crucial to reduce mortality risks. In classifying patients into severity categories, the more important classification errors are under-diagnosis, in which patients are misclassified into less severe categories and thus receive insufficient medical care. The Neyman-Pearson classification paradigm has been developed to prioritize the designated type of error. Current NP procedures are either for binary classification or do not provide high probability controls on the prioritized errors in multi-class classification. We propose a hierarchical NP framework and an umbrella algorithm that generally adapts to popular classification methods and controls the under-diagnosis errors with high probability. Beyond COVID-19 severity classification, the H-NP algorithm generally applies to multi-class classification problems, where classes have a priority order.

11:45-12:00 PM

Novel Neural Expectation-Maximization Algorithm for Semi-Competing Risk Prediction

Stephen Salerno, University of Michigan, Department of Biostatistics

Survival processes of patients with lung cancer often involve a non-terminal event, such as disease progression, and a terminal event, such as death, which form a semi-competing risk relationship. By semi-competing, we mean that the occurrence of the non-terminal event is subject to the terminal event, but not vice-versa. We propose a novel neural expectation-maximization algorithm for predicting semi-competing risk outcomes based on the illness-death model.

This approach allows us to predict patient-specific hazards for transitions between disease states, where effects of potential risk factors are estimated through a multi-task deep neural network, with hazard-specific sub-architectures, and a patient?s baseline hazard functions are estimated non-parametrically using the expectation-maximization algorithm. As deep learning can recover non-linear risk scores, we test our method by simulating risk surfaces of varying complexity. We apply our method to the Boston Lung Cancer Study, where we investigate the impact of clinical and genetic predictors on disease progression and mortality.

108. CONTRIBUTED PAPERS: INFECTIOUS DISEASES, ENVIROMENTAL EXPOSURES, AND PUBLIC HEALTH APPLICATIONS

Chair: Zhu Wang, The University of Tennessee Health Science Center

10:30-10:45 AM

A Scaler-on-Quantile-Function Approach for Estimating Short-term Health Effects of Environmental Exposures

Yuzi Zhang, Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University

Time-series analysis is widely used for estimating short-term health effects of environmental exposure by linking aggregate outcomes to exposures that are available at increasingly fine-spatial resolutions. However, areal averages are typically used to derive population-level exposure, which may not fully capture spatial variation and individual heterogeneity in exposures. We describe a general modeling approach to incorporate exposure heterogeneity via exposure quantile functions. Furthermore, by viewing exposure quantile function as a functional covariate, our approach provides additional flexibility in characterizing associations at different quantile levels. We apply the proposed approach to an analysis of air pollution and emergency department (ED) visits in Atlanta over 4 years. The analysis utilizes personal exposure to four traffic-related air pollutants simulated from the Stochastic Human Exposure and Dose Simulator. Our more nuanced analyses find that effects of carbon monoxide on respiratory and cardiovascular disease ED visits are more pronounced with changes in lower quantiles of the population's exposure.

10:45-11:00 AM

Robust Privacy-Preserving Models for Cluster-Level Confounding in Healthcare Provider Evaluations

Nicholas Hartman, Department of Biostatistics, University of Michigan-Ann Arbor

Policymakers routinely assess healthcare providers to ensure that they deliver adequate care, and adjustment models are used to control for factors beyond the providers' control. In practice, some confounding factors are defined at the provider level, or at higher cluster levels such as geographic regions. Considering that national healthcare datasets are extremely large, often contain outlying providers, and are not easily shared due to patient privacy concerns, conventional models are not well-suited to estimate cluster-level confounding effects. To address these limitations, we derive a model that only depends on public summary statistics, which are available to stakeholders, and that leverages individualized empirical null methodology to explicitly model outliers. We then develop a Pseudo-Bayesian method to flag low-quality care, while adjusting for observed and unobserved confounders. Simulations show that our estimates are robust and accurate, and the proposed flagging method has better Frequentist properties than existing approaches. We apply these methods to assess transplant centers while controlling for geographic disparities in donor organ availability.

11:15-11:30 AM

A Continuous-Time Dynamic Factor Model for Intensive Longitudinal Data Arising from Mobile Health Studies

Madeline Abbott, Department of Biostatistics, University of Michigan

Intensive longitudinal data (ILD) collected in mobile health (mHealth) studies contain rich information on multiple outcomes measured frequently over time and have the potential to capture short- and long-term dynamics. Motivated by a mHealth study of smoking cessation in which participants report the intensity of many emotions multiple times per day, we propose a dynamic factor model that summarizes the ILD as a low-dimensional, interpretable latent process. The model consists of two submodels: (i) a measurement submodel--a factor model--that summarizes the longitudinal outcome as lower-dimensional latent variables and (ii) a structural submodel--an Ornstein-Uhlenbeck process--that captures the temporal dynamics of the multivariate latent process in continuous time. We derive a closed-form likelihood for the marginal distribution of the outcome and propose a block coordinate descent algorithm for estimation. We apply our method to the mHealth data to summarize the dynamics of 18 emotions as two latent processes, which are interpreted by behavioral scientists as the psychological constructs of positive and negative affect and are key in understanding vulnerability to smoking.

11:30-11:45 AM

Circadian Blood Pressure Dysregulation in Children with Obstructive Sleep Apnea

Md Tareq Ferdous Khan, University of Cincinnati

The objective of this study was to examine 24-hour circadian blood pressure (BP) rhythms in children with obstructive sleep apnea (OSA) compared to healthy controls. We used a shape-invariant model to the 24-h BP monitoring data of 219 aged- and gender-matched children (117: controls, 52: mild OSA, and 50: moderate-to-severe OSA (MS-OSA)) to compare circadian BP patterns between the groups. It revealed that time arrived at peak velocity (TAPV) for SBP in the evening, DBP in the morning and evening, and mid-day velocity nadirs for both SBP and DBP were significantly earlier in MS-OSA group than in controls. Timing of SBP and DBP peaks and nadirs were also significantly earlier for the severe group. Similarly, the MS-OSA group reached most BP values significantly earlier than controls. Moreover, SBP was elevated in MS-OSA group in the evening (6-9 PM), and lower than controls at 12 AM. DBP was higher in MS-OSA group from 7 AM to 12 PM, and lower than controls at 12AM. The MS-OSA group was prone to non-dipping compared to controls. The findings of dysregulated circadian BP rhythms in children with MS-OSA may provide more insights for better management of MS-OSA children.

11:45-12:00 PM

Integrating Summary Information from Many External Studies with Population Heterogeneity and a Study of COVID-19 Pandemic Impact on Mental Health of People with Bipolar Disorder

Yuqi Zhai, Department of Biostatistics at the University of Michigan

For integrating summary information from external studies to improve internal model fitting when study population heterogeneity exists, Zhai and Han (2022) proposed a penalized constrained maximum likelihood (PCML) method that can simultaneously select and incorporate only the useful part of the external information. Their work, however, only considered cases where the number of external studies is small. Motivated by a study of the COVID-19 pandemic impact on mental health of people with bipolar disorder (BD), we extend the PCML method by allowing the number of external studies to increase with the sample size of the internal study. Within this more general framework, we re-establish the asymptotic properties of the PCML estimator, including external information selection consistency and oracle efficiency, and carry out comprehensive simulation studies. The method is then applied to the motivating study to integrate useful external information from many existing mental health studies that are not for people with BD to improve internal study effect estimation. Integrating external information helps to reveal interesting time trends from pre-pandemic to pandemic periods.

# # #

**Indicates a 2023 Distinguished Student Paper Award winner*