***POSTER ABSTRACTS – STUDENTS***

**1. Identification of Significant Interactions using Bayesian Additive Regression Trees**

**Presenter: Joshua Marvald, University of Rochester Medical Center**

**Co-authors: Tanzy Love, University of Rochester**

In the age of personalized medicine, understanding the complex, potentially non-linear, interactions amongst an ever-growing pool of covariates remains a challenge. While parametric methods can identify interactions if distributional assumptions are met and functional forms are known, few methods exist that work to identify interactions without prior knowledge of the covariates involved. Some non-parametric or machine learning algorithms provide evidence of interactions in relation to other possible interactions, but most stop short of selection. Leveraging Bayesian Additive Regression Trees (BART), the presented method uses posterior tree co-inclusion probabilities in combination with null run thresholding to select interactions from a prediction model. Advantages of this method include few distributional assumptions, ease of scalability to high-dimensional data, ability to include prior scientific knowledge, and selection of important variable interactions. In addition, BARTâ€™s flexible sum of trees structure allows for high-order or non-linear interactions whose form would be inaccessible a priori.

**2. Bayesian Interval Estimation for Predictive Values from Continuous Diagnostic Tests in Case-Control Studies**

**Presenter: Jamie Roberman, Baylor University**

**Co-author: James Stamey PhD and Melinda Holt PhD, Baylor University**

Diagnostic tests based on continuous biomarkers often depend on an estimated cut-off point that dichotomizes the biomarkers as diseased and non-diseased. Positive predictive and negative predictive values (PPV and NPV) can be used to evaluate the diagnostic accuracy of such tests. PPV and NPV are functions of test accuracy and disease prevalence. Previous studies providing a joint estimation of the predictive values assume a fixed prevalence. However, prevalence is often unknown. Using a Bayesian prior allows for uncertainty in estimating the population prevalence. We use the Alzheimer's Disease Neuroimaging Initiative (ADNI) data set to compare our new approach to an existing method and to demonstrate how different prevalence estimates affect the interval estimation. Finally, simulation results indicate that increased levels of uncertainty result in wider intervals that are more robust to deviations from the estimated prevalence.

**3. Bayesian Data Augmentation for Recurrent Events with Intermittent Assessment in Overlapping Intervals**

**Presenter: Xin Liu, The Ohio State University**

**Co-authors: Patrick Schnell, The Ohio State University**

Existing methods analyzing intermittently assessed interval-censored recurrent events assume disjoint assessment intervals (interval count data). This is due to a focus on prospective studies with controlled assessment times. Electronic medical records (EMRs) databases have much richer information but the measurements for recurrent events aren't made as interval count data, e.g., subjects are asked have they fallen within the last 90 days at uncontrolled appointment times. This can result in non-contiguous or overlapping assessment intervals and censored event counts for the assessments. A Bayesian data augmentation method is proposed to utilize the complicated assessments for the recurrent events. In a Gibbs sampler, the exact event times are imputed by rejecting simulations of non-homogeneous Poisson process times that are incompatible with the assessments. Three optimizations have been implemented to speed up the rejection sampling process for large EMR datasets: truncated Poisson process simulation, independent sampling by partitioning, and sequential sampling. We illustrate our method by application to the EMR dataset on falls among breast cancer patients and survivors. While Poisson process assumes independent increments in events, renewal process, which accounts for serial correlations between events, can also be used for imputation and modeling. In this application, we compared results from Poisson process and renewal process.

**4. Detecting Influential Subjects in Intensive Longitudinal Data Using Mixed Effects Location-Scale Models**

**Presenter: Xingruo (Summer) Zhang, The University of Chicago**

**Co-authors: Donald Hedeker, The University of Chicago**

Collection of intensive longitudinal health outcomes allows joint modeling of their mean (location) and variability (scale). Focusing on the location of the outcome, measures to detect influential subjects in longitudinal data using regular mixed-effects regression models (MRMs) have been widely discussed. However, no existing approach enables the detection of subjects that heavily influence the scale of the outcome. We propose applying mixed-effects location scale (MELS) modeling combined with commonly used influence measures such as Cookâ€™s distances and DFBETAS to fill in this gap. In this poster, we provide a framework for researchers to follow when trying to detect influential subjects for both the scale and location of the outcome. The framework allows a detailed examination of each subjectâ€™s influence on model fit as well as point estimates and precision of coefficients in different components of a MELS model. A practical example based on data from a health behavior study, together with simulated examples, are given to illustrate the benefits of using MELS models to detect influential subjects in intensive longitudinal data compared with MRMs.

**5. Two-stage Bayesian analysis for Time series data**

**Presenter: Richard Ngaya, University of Nebraska Medical Center**

**Co-author: Dr. Yeongjin Gwon, Dr. Jesse Bell, University of Nebraska Medical Center**

A Two-stage Bayesian Approaches for Analyzing Time series data: Applications to Drought and Health Impacts study in the United States ? Richard Ngaya, Dr. Yeongjin Gwon, and Dr. Jesse Bell Abstract: Drought affects more people than any other natural disaster that are complex and least understood [15,16]. A suitable and more accurate prediction model is needed to study the complexity in studying drought and its relationship to human mortalities related to cardiovascular and respiratory. Two-stage modeling provides an opportunity to capture the complex between-study heterogeneity and spatial-temporal structure in the data. A Bayesian approach in the second study provides more powerful opportunity to solve the computational challenges involving poor mixing from the generated MCMC and model convergence issues related to count and time-series data. In this poster, we will discuss a variety of model that suits the analysis of time-series data from different locations (counties in Nebraska), collected over 19 years on total, cardiovascular, and respiratory mortalities related to drought exposures. The first-stage model (quasi-Poisson regression) model will be presented followed by its estimated drought effect from each county presented through maps (made through ArcGIS Pro software). Bayesian analysis with two approaches (Fixed effect metal-analysis, and Random-effect meta-analysis) will be presented as part of the second-stage analysis for the data based on the estimates obtained from the first-stage analysis. In the second stage analysis, the overall effect of drought, and itâ€™s between-study heterogeneity will presented to help understand the drought exposure effect to human health through total, cardiovascular and respiratory mortalities. References. 1. Ma, L., Yan, X., & Qiao, W. (2014). A Quasi-Poisson approach on modeling accident hazard index for urban road segments. Discrete Dynamics in Nature and Society, 2014, 1-8. 2. Lynch, K. M., Lyles, R. H., Waller, L.

**6. A Bayesian Nonparametric Model for External Data with Application to Clinical Trials**

**Presenter: Dehua Bi, The University of Chicago**

**Co-authors: Yuan Ji , The University of Chicago**

We consider a new Bayesian nonparametric model that borrows information across external data sets with an aim enhance the efficiency of an ongoing clinical trial. The model, named Plaid Atoms Model (PAM), induces a dependence clustering structure across grouped data sets. Each atom, i.e., cluster, represents a homogeneous subpopulation of patients that are believed to respond to a treatment or control arm similarly. PAM automatically infers commons atoms as well as unique atoms for each data set so that information from common atoms can be shared. In this fashion, the efficiency of a clinical trial may be improved if the patient samples in the trial share common clusters with external data. Applications of PAM include augmenting or synthesizing a control arm using external data. Examples will be provided.

**7. Statistical Inference for Response-Adaptive Randomization Designs: A Comparative Study**

**Presenter: Xue Yang, University of Pittsburgh**

**Co-authors: Yu Cheng, Abdus S. Wahed, University of Pittsburgh**

Response-adaptive randomization (RAR) aims to achieve greater patient benefit by assigning more patients to more efficacious treatments maintaining the validity of treatment comparisons. While numerous RAR designs have been developed in the literature, sample average treatment effect (ATE) remains to be the basis of inference for causal treatment effect. In this article, we propose some alternative estimators of causal treatment effect for RAR designs based on inverse-probability-weighting (IPW) and compare them to the sample ATE analytically based on bias, variance, type I error and statistical power. We conducted extensive simulation studies to assess the operating characteristics of these estimators. Results from the simulation studies show that when implemented and analyzed correctly, RAR can treat substantial proportion of the patients using the better treatment minimally sacrificing the power compared to a standard allocation design. However, when both the sample size and the response rate are small, RAR fails to control the type I error. The analytical and simulation results also indicate that the IPW-based estimators are consistent and have smaller bias, and are more efficient compared to the sample ATE, especially in the presence of covariates. Moreover, the methods based on log relative risk and log odds ratio control the type I error better than the sample ATE-based method.

**8. Fusion Learning for Heterogeneous Treatment Effect Estimation using Distributed Data**

**Presenter: Jinhong Li, University of Pittsburgh**

**Co-authors: Lu Tang, University of Pittsburgh**

When estimating conditional average treatment effect (CATE) in data-driven precision medicine, data integration is often undertaken to achieve a larger sample size and better statistical efficiency. Since data from different sources may be inherently heterogeneous, challenges arise in how to optimally and securely combine information while accounting for between-study heterogeneity. In the meanwhile, privacy concern may prevent the sharing of individual-participant data. In this paper, we propose a two-step doubly robust algorithm: we first estimate the marginal effects and propensity scores in each distributed data to formulate an objective function for estimating CATE; we then aggregate study-level data from model estimates to optimize the objective function. Specifically, in the second step, we leverage fused lasso to capture heterogeneity among different data sources, and confidence distribution to preserve data privacy and avoid the sharing of individual-participant data. The performance of this approach is evaluated by simulation studies and a real-world study of the causal effect of oxygen saturation target on survival rates for multi-hospital ICU patients.

**9. Identifying Regulators of Intestinal Regional Identity using Time Series analysis of Pseudo-Continuous RNA-seq Data**

**Presenter: Yuqi Zhang, Washington University in St. Louis**

**Co-author: Anna Bass, Brian Muegge, Washington University in St. Louis**

The intestine is a linear epithelial tube, but the anatomic simplicity of this organ belies its functional and cellular complexity. Suites of genes controlling digestion, defense, and endocrine signaling are expressed in characteristic distributions along the proximal to distal axis. The molecular regulators that establish and maintain the proximal to distal identity are not well understood, but regionalization is important for diseases including inflammatory bowel disease and obesity. We hypothesized that regional identity in the intestine is established by competing gradients of gene expression along the longitudinal axis. To identify the regulators of regional identity, we surveyed intestinal gene expression along the length of the mouse gut at high anatomic resolution. We used a novel, high throughput bulk 3â€™ mRNA-sequencing method to perform RNA-sequencing on samples obtained every ~1 cm along the length of the intestine from the stomach to the rectum. Preprocessing and quality control was performed in R using edgeR-Limma. We applied supervised and unsupervised clustering approaches to identify suites of genes that varied across the length of the gut. Unsupervised clustering demonstrated a sharp distinction between samples from the small and large intestine. Potential regulators of this boundary include many well characterized members of the GATA and HOX transcription factor families. We also identified many highly specific but poorly characterized transcription factors in each region, which we are testing functionally. Within the small intestine, the expression of many genes fell on a pseudo-continuous distribution from the proximal to distal sections. We adapted methods from time-series analysis to define these distributions and identified suites of co-occurring genes that might establish distal vs proximal identity. Together, our results have identified transcriptional regulatory pathways that might contribute to establishing intestinal regional identity.

**10. A Generalized Influence Maximization Problem**

**Presenter: Sumit Kumar Kar, University of North Carolina at Chapel Hill**

**Co-authors: Nilay Tanik Argon, Shankar Bhamidi, Serhan Ziya, University of North Carolina at Chapel Hill**

Influence maximization problem (IMP) is a popular topic in social networks with several applications in viral marketing and epidemiology. An important application is in epidemic outbreaks like COVID-19 which start from an initial infected set of people and then spread as a viral contagious process, where the infection spread is modeled by some standard diffusion model in which infected nodes infect other nodes without external intervention. The goal of IMP is Targeted Immunization under a budget constraint, i.e., to choose an optimal set of people within a given budget so that immunizing this set maximizes the expected number of people who would have been infected without this immunization. Kempe et al. showed that IMP is NP-hard under the Triggering model, but a greedy algorithm-based approach can provide a (1-1/e)-approximation guarantee compared to the optimal solution. In our work, we consider a much more general problem where the infection spread is considered until a given finite or infinite time. Moreover, we generalize the concept of maximizing the expected number of people protected from infection by maximizing the expected reward obtained from these people. The reward is modeled by a set function that captures importance of individual people by assigning weights, is not necessarily additive over the people, and depends on the times at which the people would have been infected in absence of immunization. Furthermore, we allow administration of multiple doses of immunization (e.g., for a vaccine) where there may be restrictions on whether a person may be available for immunization and if they are, they may have their own capacity for the number of doses they may be given. Our generalized Triggering model also recognizes that more doses provide stronger immunity and reduce the chance of spreading infection to others. We formulated a greedy algorithm that achieves a (1-1/e)-approximation guarantee compared to the optimal solution of this generalized IMP.

**11. Bayesian Historical Borrowing using Relative Belief Ratios**

**Presenter: Nghia Nguyen, Baylor University**

**Co-authors: John Seaman, Baylor University**

A powerful advantage the Bayesian paradigm has over its classical frequentist counterpart is the ability to borrow information from past studies to help empower a current study with small sample size instead of just combining them into a larger sample in the latter case. In particular, the methodology of historical borrowing in the one-sample proportion case is explored in details using relative belief ratios and its various ingredient-checking measures to see how much of the historical data one can borrow to use empower the current study without incurring prior-data conflict or inducing too much bias, either against or in favor of the null, in the result. Various choices of the beta mixture priors were chosen in the simulation study to learn about the method's operating characteristics, so informed decisions can be made in accordance with the FDA guidelines, but the results can be extended to cases where power priors and commensurate priors are used. Applications to clinical trials, especially pediatric extrapolation, are discussed.

**12. Outlier Resistant Inference for Conditional Average Treatment Effect**

**Presenter: Ran Mo, IUPUI**

**Co-author: Honglang Wang, 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 weight- ing (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 purpose. The finite sample performance of the proposed estimator is evalu- ated 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.