Biostatistics Student Research Symposium (BSRS) 2022

Abstracts

VCU Biostatistics
School of Medicine

Title: *Cluster Competing Risks with Informative Cluster Size: A Fast and Scalable Approach*
Presenter: Edem Defor
Advisor: Dr. Dipankar Bandyopadhyay
Abstract:

Large-scale electronic health records are becoming increasingly accessible to biomedical researchers. The clustered competing risk (CCR) model assesses the marginal effects of covariates on the cumulative incidence function. However, this does not account for informative cluster sizes, for example, in center-volume outcomes. CCR implementations are also not computationally scalable for large-scale data. Our research focuses on extending the Zhao et al. CCR model (2012) to account for situations where the cluster size is related to the outcome of interest. We also demonstrate an impressive computational efficiency by extending and implementing the Kawaguchi et al. (2020) forward-backward scan algorithm. We propose applications of these procedures in an R package. Data on kidney transplants between 1999-2015 from the United Network of Organ Sharing is utilized.

Title: *Marginal Feature Screening and FDR control for the Ultrahigh Dimensional Right Censored Survival Outcomes.*
Presenter: Atika Farzana Urmi
Advisor: Dr. Dipankar Bandyopadhyay
Abstract:

In many large-scale biomedical studies, often time the primary outcome is the time to an event subject to right censoring. The existence of censoring along with ultra-high dimension and outliers or heavy tailed predictors bring great challenges for feature screening. In this study, we propose a two-step variable screening method along with False Discovery Rate (FDR) control under a pre-specified level. The proposed screening method is based on the Expected Conditional Characteristics Function Based Independence Criterion (ECCFIC), which measures the dependence between two random vectors. The ECCFIC does not require any specific model-assumption and can be applied on both continuous and categorical outcome. The FDR control is done by using knockoff feature-a fake variable considering as a control for each original covariate selected in the screening stage. We apply our method to analyze a head and neck cancer (HNSCC)study on detecting prognostic gene and clinical features for cancer patients' survival. Existing screening methods are compared using cross validated AUC values.

Title: *Reassessing Replication: Addressing the Replication Crisis from a Statistical Perspective*
Presenter: Alicia Richards
Advisor: Dr. Robert A. Perera
Abstract:

Introduction: Replication, defined as obtaining consistent results using newly collected data following the original studies population and protocol, is used to assess the validity and reliability of research findings. Recently, the lack of successful replication of published studies has led to concern of a replication crisis. Scholars have offered potential reasons for the low replication rates including flawed statistical methods to assess replications. Currently, the common methods to assess replication dichotomize replication success and do not account for study limitations. Thus, this study aims to build a metric that assesses replication continuously while addressing underpowered studies and publication bias.

Methods: The Reproducibility Project's data and simulated data, were applied to a novel metric which uses equivalence tests to assesses replication success. Equivalence margins were based on the effect size (ES) of the original study and the ES differences between the studies. The replicated studies ES interval was then used to determine the studies ability to replicate. Additionally, the equivalence replication success rate results were compared to previous metrics.

Results: For the equivalence metric, a studies ability to replicate was higher when the ES difference between the original and replicated studies was smaller. However, the sample size of the original study, the design of the replicated study, and the studies power level highly impacted a studies ability to replicate.

Conclusion: Using equivalence studies to assess replication allows replication success to fall on a continuous scale providing more information about the studies ability to replicate while controlling for design limitations.

Title: *An Adaptive Method for Covariate Balancing in Block Randomized Clinical Trials*
Presenter: Rasnick
Advisor: Dr. Leroy Thacker
Abstract:

Introduction: Small sample sizes are prone to randomization bias and can present challenges when conducting clinical trials. Randomization bias occurs when the treatment and control groups differ on a key demographic (confounder); confounders when not properly adjusted can increase unexplained variability, thus making the results from the study less precise. Current methods in trial design that limit randomization bias include stratified block design, minimization, and propensity scores.

Methods: The binomial and multinominal distributions were used to calculate the expected final number of participants with a certain categorical characteristic in each of the treatment and control groups, separately, using the information from participants already in the study, an assumed prevalence of the characteristics, and the known final sample size of the trial. The difference between these expectations is the expected imbalance, and if this value was greater than a threshold, the original allocation was adapted to reduce the predicted future imbalance. The binary case is reviewed in this talk and we propose three possible ways for a multivariate case.

Results: Previously we showed for the binary case, we were able to maintain Type 1 error while decreasing the variability of the expected imbalance throughout each trial when the confounder had no effect on the trial compared to a simple block design.

Discussion: This method is inventive in that it adapts for what allocation is expected to happen instead of what has already been observed. We intend to extend our adaptive trial methodology to adjust for many variables, both continuous and categorical, in hopes to improve current standards and adjust for unknown prevalence of the characteristic.

Title: *Variability in the Complier Average Causal Effect of e-Assist on a Binary Outcome by a Shared Random Effects Model*
Presenter: Xinxin Sun
Advisor: Dr. Yongyun Shin
Abstract:

e-Assist is a decision aid program to help eligible patients complete colorectal cancer screening (CRCS); they are randomized to e-Assist or control within physicians (PH). However, the effect of e-Assist may vary randomly over heterogeneous PH and compliance C to treatment assignment is imperfect. Because those assigned to control cannot access e-Assist, one assigned to e-Assist is an "e-Assist complier" if the patient takes the assignment or ``never taker'' otherwise; one assigned to control, however, could be a "control complier" if she would have taken e-Assist or "never taker" otherwise under the alternative assignment. Assuming C missing at random, we jointly model CRCS and C for the random, PH-specific CRCS rates of the C groups; the difference between e-Assist and control compliers is a complier average causal effect. We compare the means, variances and covariances of random rates to learn who compliers are. We integrate random effects out by adaptive Gauss Hermite quadrature (AGHQ) to compute the likelihood. To deal with the intensive computation by AGHQ and highly collinear random effects, we estimate the joint mixed model with shared random effects by maximum likelihood.

Keywords: One-sided compliance, missing at random, adaptive Gauss Hermite quadrature, the EM algorithm; Newton Raphson, maximum likelihood

Title: *Estimating mixture effects and cumulative spatial risk over time simultaneously using a Bayesian index low-rank kriging multiple membership model*
Presenter: Joseph Boyle
Advisor: Dr. David C. Wheeler
Abstract:

The exposome is an ideal in public health research that posits that individuals experience risk for adverse health outcomes from a wide variety of sources over their life course. There have been increases in data collection in the various components of the exposome, but novel statistical methods are needed that capture multiple dimensions of risk at once. We introduce a Bayesian index low-rank kriging (LRK) multiple membership model (MMM) to simultaneously estimate the health effects of one or more groups of exposures, the relative importance of exposure components, and cumulative spatial risk over time using residential histories. The model employs an MMM to consider all residential locations for subjects weighted by duration and LRK to increase computational efficiency. We demonstrate the performance of the Bayesian index LRK-MMM through a simulation study, showing that the model accurately and consistently estimates the health effects of one or several group indices and has high power to identify a region of elevated spatial risk due to unmeasured environmental exposures. Finally, we apply our model to data from a multi-center case-control study of non-Hodgkin lymphoma (NHL), finding a significant positive association between one index of pesticides and risk for NHL in Iowa. Additionally, we find an area of significantly elevated spatial risk for NHL in Los Angeles. In conclusion, our Bayesian index LRK-MMM represents a step forward towards bringing the ideals of the exposome into practice for environmental risk analyses.

Title: *Early Termination in Phase II Clinical Trial Admissible Designs Using Decreasingly Informative Priors for Two-parameter Models*
Presenter: Chen Wang
Advisor: Dr. Roy T. Sabo
Abstract:

We extend Neuenschwander etc. approach for calculating the prior effective sample size (ESS) from one-parameter to two-parameter models. We then functionalize the ESS in the Bayesian decreasingly informative priors (DIP) used for early termination in Phase II clinical trial designs. Simulation studies are conducted to compare the ESS-DIP approach with the standard Bayesian approach (Thall & Simon's approach) to identify the smallest possible sample size among admissible designs, which is defined as having at least 80% power and no greater than 5% type I error rate. We identified admissible designs by searching through all possible values of total sample sizes (N) and decision criteria (pf, ps) to find the smallest total sample size (N) under the admissible power and type I error. Two examples are presented: Normal distribution with unknown mean and variance, and Weibull distribution with unknown scale and shape parameters. Simulation studies in Normal distribution show: (1) the DIP approach requires comparable or fewer patients when admissible designs are achieved; (2) the DIP approach yields similar power and better-controlled type I error with comparable or fewer patients than Thall and Simon's Bayesian approach. The results of the DIP approach help to pre-determine termination rules for total sample size in the way that is not based on any historical or optimistic prior.

Title: *A Comparison of Gene Co-regulation Pattern Analysis Methods with Multi-group RNA-seq Data*
Presenter: Khoa Huynh
Advisor: Dr. Jinze Liu
Abstract:

RNA-sequencing (RNA-seq) measures the quantity and diversity of RNAs in biological samples leveraging next-generation sequencing technology. In recent years, decreasing cost of RNA-seq has made it increasingly popular to design multigroup comparison and/or time series experiments. While powerful, these experimental designs lead to new challenges in data analysis. Existing methods are often based on aggregating two-group comparisons to detect differential gene expression. However, such an approach failed to render a complete picture across all groups. Alternatively, high dimensional data analysis can detect gene regulation patterns using entire datasets. In this study, we surveyed three recently developed methods: bigPint, weighted gene correlation network analysis (WGCNA), and Bayes mixture modeling approach (EBSeq-HMM) to identify gene co-regulation pattern analysis in multi-group RNA-seq. Both the bigPint and WGCNA utilized hierarchical clustering to classify gene co-regulation patterns. But bigPint requires pairwise differential gene expression while WGCNA calculates the pairwise correlation between genes in RNA-seq. EBSeq-HMM uses an auto-regressive hidden Markov model to implement gene expression across time conditions. We applied and evaluated those methods to a study involving 60 RNA-seq samples captured in nematode C. elegans to identify gene patterns regulated by omega-3 polyunsaturated fatty acid eicosatetraenoic acid (EPA) over three-time points.

Title: *Empirical Grouping of Chemical Mixture Components with Bayesian Group Index Regression*

Presenter: Matthew Carli
Advisor: Dr. David C. Wheeler

Abstract:

An area of growing scientific interest is the study of environmental chemical mixtures and their effect on human health. One method developed for this purpose is Bayesian group index regression, where a mixture of chemicals is divided into a number of discrete indices which are then regressed on an outcome of interest. Two questions that naturally arise from this modelling strategy are how many indices should be formed, and which chemicals should be grouped together. In previous applications of the Bayesian group index model, group number and composition were determined by similarities in chemical structure or usage (e.g. metals with metals, pesticides with pesticides, etc.). We propose to group chemicals empirically using the Robust Principle Components Analysis (RPCA) algorithm and subsequently use these indices in the group index model. We compare the performance of RPCA with PCA and other PCA variants. To evaluate our proposed method, we conduct simulation studies with varying numbers of true groups, strengths of association, and noise. Our results indicate that RPCA is superior to other implemented PCA variants in assigning chemicals to the correct number of groups. Further, after they are used in Bayesian group index regression the indices formed by RPCA lead to higher power for exposure effects as well as higher sensitivity and specificity for identifying the relative importance of individual chemicals. In conclusion, the use of RPCA provides an empirical basis for grouping chemicals and aids the model-fitting process for Bayesian group index regression.

Title: *An Integrated Multiple Adaptive Design for Covariate-Adjusted Response-Adaptive Randomization and Sample Size Re-estimation in a 2x2 ANOVA Setting*
Presenter: Serenity Budd
Advisor: Dr. Robert A. Perera
Abstract:

Adaptive randomized clinical trial designs are growing in popularity due to their increased efficiency and ethical benefit. Although adaptive designs are more frequently being implemented, usually, only one adaptive feature is incorporated into the design of the trial. This research employs integrated multiple adaptive design to optimize the multiple objectives of a covariate-adjusted response-adaptive randomization and sample size re-estimation for a 2x2 ANOVA. Specifically, the proposed method is designed with the intention to power the tests of simple effects of treatment for each level of the covariate. The goal is to develop a method that allocates participants to the more effective treatment group given their covariate value and minimizes sample size, while at the same time maintaining standard levels of type one error and power. Our proposed method is evaluated using simulation studies under various scenarios that modify the strength of the treatment effect, presence or absence of an interaction effect, and the balance of covariate values in the sample. The successful application of the method should yield a more efficient clinical trial design that allows for increased individualization of treatment randomization.

Title: *Investigating the Moderating Effect of Small Intestine Bacterial Growth (SIBO) on the Effectiveness in a Two-Factor Study of Antimicrobial and Nicotinamide Treatment on Growth in Children.*
Presenter: Rasha Alsaadawi
Advisor: Dr. Roy T. Sabo
Abstract:

Stunted growth in infancy due to environmental enteric dysfunction is a persistent problem in developing countries. While there is potential for antibiotic treatment to help reduce or prevent this phenomenon, such treatments have remained elusive. In this randomized, double-blind, placebo-controlled, two-factor trial, we aim to identify levels of small intestine bacterial growth (SIBO) for which the two-factor intervention of antimicrobial treatment and nicotinamide treatment is effective at increasing growth in Tanzanian children. We first determine optimal SIBO cut-off levels to split the data into two sub-groups of patients: those with low SIBO and those with high SIBO. The primary outcome of interest in the study is the change in body length Z-score from 0 to 18 months, while four MDAT scores are included as secondary outcomes: Fine Motor Score, Gross Motor Score, Social Development Score, and Language Score. For each outcome, a two-factor ANOVA model is fit in each SIBO sub-group to test the interaction effect and the main effects of the antimicrobial treatment and nicotinamide treatment. The models are fit again with adjustment for some covariates like gender, WAMI score, maternal height, and water treatment indicator. We use LASSO regression to select model covariates that we are adjusting for. Our results indicate that portions of the two treatments were effective in patients with high SIBO scores, with some effectiveness remaining after adjustment. These results provide evidence that the potential to reduce stunted growth is likely tied to treatments that also promote increased intestinal biodiversity.

Title: *Quantifying major sources of variability in microbiome sequencing*
Presenter: Adam Funk
Advisor: Dr. Ekaterina Smirnova
Abstract:

The composition of microbiome in and on human body has been associated with multiple diseases including inflammatory bowel disease, type 2 diabetes, hepatitis, and many others. Given the high prospects of using microbiome data to aid in patient health decisions there is an increasing need for reliable data processing methods that can be reproduced across multiple studies. However, current within-lab microbiome data processing methods result in large differences in the type and abundance of observed microbial organisms. These challenges further prevent horizontal integration of microbiome data collected by multiple independent studies. Thus, it is of interest to understand the aspects of the complex process of sequencing microbiome data with a goal of adjusting for the lab differences in the integrated data analysis. In our study, we concentrate on understanding if there is a controllable aspect of the sequencing process that induces more variability than one would expect from uncontrollable error to explain where this difference arises. We use data collected from the Microbiome Quality Control Project (MBQC) where 16 sample handling laboratories processed identical stool samples aliquots using in-house protocols. We evaluate differences in each step of the sample handling and processing protocols by comparing within sample microbiome summary measures, known as alpha diversity, for identical samples. Four commonly used alpha diversity measures are used: Shannon index, Chao1 index, Inverse Simpson index, and number of observed species. To quantify the relative contribution of each step of the sequencing protocol on alpha diversity, ANOVA percentage of variability and ranked contribution per degree of freedom within each scoring method and aliquot is used. Results indicate that pre-extracted samples and polymerase chain reaction (PCR) kit manufacturer explain significant variability in several aliquot scores, but their impact varies from sample to sample.

Title: *Linear Mixed Models with Cluster-Level Interactive Terms (MAR): Multiple Imputation by the Gibbs Sampler via Latent Cluster Means*
Presenter: Dongho Shin
Advisor: Dr. Yongyun Shin
Abstract:

In a two-level hierarchical linear model, cluster-level covariates are partially observed and continuous. A pervasive missing data analysis is to estimate the distribution of the outcome and covariates as the joint imputation model to impute missing data. When the covariates are interactive, however, the joint distribution will not be compatible with the hierarchical model. In this research, we find our joint imputation model for the outcome and covariates, compatible with the hierarchical model, that specifies the joint distribution of the latent cluster-level mean outcome and covariates. Via the joint imputation model, we estimate the hierarchical model efficiently by the Gibbs sampler. We apply this approach to real data and assess our estimators by simulation.

Title: *Topological Data Analysis for Tree Based Predictive Model*
Presenter: My Nguyen
Advisor: Dr. Nitai Mukhopadhyay
Abstract:

Clustering analysis is a fundamental topic in machine learning that is widely applied in market and health research, including cancer and genomics studies. One of the primary methods for cluster analysis is based on regression tree algorithm, which performs very well. A recent method to build a hierarchical clustering is level set cluster tree based on topological data analysis (TDA, Wasserman), a technique to find the structure of data by distance as opposed to likelihood. Our hypothesis of interest is to explore if the application of TDA would improve the predictive performance of the model by comparing mean square error between the tree-based methods and TDA based cluster trees. However, the existing software only computes a density estimator and returns the corresponding cluster tree of super-level sets, but not being able to conduct prediction from that level set tree. Hence, we apply TDA level set clustering to build a predictive model for the 2D medical cost personal dataset by utilizing computation of Euclidean distance and then classifying data points into clusters with minimum distance to those points. We present the simulation study for synthetic datasets generated from the observed data with different structures of clusters by varying their clusters' means and then evaluate the tree-based model and TDA cluster tree within each of these sampling datasets. We find that the prediction performance of the tree-based model and TDA level set tree are comparable in most of the study cases. We will discuss the limitations of the algorithm and outline future analyses toward exploring the TDA tree in the higher dimensional data.

Database reference: Lantz, Brett. "Machine Learning with R". Packt Publishing Ltd, 2013.
Medical Expenses.

Title*: Development and Applications of Propensity Score Matching with Time-Dependent Covariates*
Presenter: Dustin Bastaich
Advisor: Dr. Bassam Dahman
Abstract:

Propensity score matching has become a popular method used to draw causal inferences by removing the selection bias and balancing treatment groups in the non-randomized observational studies. Methods have been developed to perform propensity score matching with time-dependent covariates to balance treatment groups based on a patient profile leading up to treatment rather than just baseline snapshots.

The time-dependent propensity scores can be generated using a Cox proportional hazard model which models time to treatment with time-varying covariates. This allows for a treated patient to be matched with a not-yet-treated control who has similar time-dependent covariates up until that treatment time. Current research has been done in this area where the hazard component of the model is used as the propensity score.

Recent research in 2021 used these methods to estimate the impact of kidney transplantation on survival. Time-dependent matching is beneficial in this area because matching on the baseline characteristics at the time of initial inclusion on the waiting list for transplantation will not account for the potential confounding information continuing to occur after wait-list inclusion.

These matching methods considering time-dependent covariates allow matching on patient profiles up until treatment which provides benefit in research areas where patient characteristics may be consistently changing. For future work we are looking to develop methods to apply this matching with more than two treatment groups. We are also interested in handling this matching method when the Cox proportional hazard assumptions are not met.

Title: *Modeling Sleep Patterns of Aging Adults in the MESA Sleep Study using Multistate Markov Chains*
Presenter: Johnathan Jacobs
Advisor: Dr. Shanshan Chen
Abstract:

Sleep hypnograms, which represent the various stages of sleep, are the categorical, intensive longitudinal outcomes from sleep quality studies. Such data can be modeled by multistate Markov chains, where the transition quantities (e.g. intensity, probability, sojourn time) between states are parameters of interest. In this project, we examined sleep hypnogram data from 1968 aging adults (aged 54 to 90) in the Multi-Ethnic Study of Atherosclerosis sleep study and fit a multistate Markov model to analyze the data, with seven states including wake to falling asleep (WB), light sleep (S1), deeper sleep (S2), deepest sleep (S3), rapid eye-movement (REM), wake after sleep onset (WASO), and final wake up (WE). Besides handling time-inhomogeneity with a time-varying phase covariate, we also assessed the associations between the hypnogram and covariates such as age group, job schedule, and self-reported sleep quality. Lastly, we visualized the transitions between states by network graphs and interpreted the transition intensity rates. Results show that both aging and job schedule have significant impacts on sleep patterns: older subjects and shift workers tend to transition more quickly to lighter sleep stages and the WASO state, with split shifts being the worst among all types of shift jobs. On average, subjects who reported having trouble falling asleep 3-5 times a week spent 2.4 more minutes in the WB state compared to those who did not report having trouble falling asleep. To conclude, multistate Markov models are useful for studying the complex network of sleep stages in hypnogram data.