Biostatistics Student Research Symposium (BSRS) 2020

Abstracts

VCU Biostatistics
School of Medicine

Title: *Understanding Algorithms to Adjusting Clinician Panel Size in Primary Care Practices*
Presenter: Martin Lavallee
Advisor: Dr. Roy T. Sabo
Abstract:

Primary care clinicians serve as the front line of patient care in a health system, ensuring day to day wellness, health education, early detection of sickness and coordination with specialty care for patients. To establish an efficient healthcare system, primary care empanelment is vital in ensuring access, continuity, and quality of care. Empanelment defines the relative burden a clinician faces through the volume of health services required by the set of persons who which they provide care, adjusting for various individual demands and requirements. Primary care practices are interested in accurately estimating panel size, ensuring that resources are efficiently and equitably distributed among their patients. Most simplistically, empanelment can be calculated using a clinician's patient count. However, since each patient has vastly differing needs, we explored how empanelment can be more accurately estimated using a vector of healthcare utilization. Examples of healthcare utilization include number of visits, number of prescriptions, number of visits to a specialists or ER, and number of communications between the clinician and patient (such as phone calls or patient portal messages). In this research, we have reviewed three algorithms for panel size calculation: first, a hybrid approach blending k-means clustering with a deterministic designation of utilization phenotypes proposed by Rajkomar et al 2016; second, a finite mixture model approach; and third, a novel geometric approach. Further work on algorithm evaluation and comparison is in planning, however for this presentation we will demonstrate the results from our production level pipeline for panel size calculation of four family medicine practices in the Richmond area.

Title: *Integrated Multiple Adaptive Clinical Trial Design Involving Sample Size Re-Estimation and Response-Adaptive Randomization for Continuous and Binary Outcomes*
Presenter: Christine Orndahl
Advisor: Dr. Robert A. Perera
Abstract:

Recent interest has emerged in the area of multiple adaptive designs, where more than one adaptive design is utilized within a single clinical trial. Usually, multiple adaptive components are performed independently and sequentially, limiting the flow of information between adaptive designs. Our proposed method integrates two adaptive designs, sample size re-estimation (SSR) and response-adaptive randomization (RAR), into a clinical trial with continuous or binary outcomes. Weighted sum multi-objective optimization is used to simultaneously optimize two objective functions: one minimizing non-ideal patient outcomes and the other minimizing the total sample size required for the trial. This allows the aims of RAR and SSR to be considered concurrently, resulting in allocation ratios and sample sizes that inform each other and are adaptively adjusted throughout the trial. Performance of our methods in simulated clinical trials is presented and compared to a variety of alternative methods, such as singly adaptive designs and sequentially adaptive designs.

Title: *A Prediction-based Replacement Algorithm for Adaptive Allocation of Severely Delayed Outcome Data via Regularization*
Presenter: Salem Rustom
Advisor: Dr. Robert Perera
Abstract:

Response Adaptive (RA) designs have been primarily developed as an ethical response to the issue of subjecting study participants to inferior treatments (by minimizing allocation to them) without compromising internal validity. However, if time-to-response is too long, then adaptive allocation cannot effectively ameliorate the issue. Nowacki et al. (2015) addressed this problem through a Surrogate-Primary (S-P) replacement algorithm in which an (earlier obtained) surrogate outcome is used in response adaptive randomization until the primary outcome becomes available to replace it. The replacement algorithm implements the Doubly Adaptive Biased Coin Design (DBCD) to estimate the target allocation ratio. We develop a replacement algorithm using earlier repeated measures of the primary outcome to predict the final outcome via Tikhonov (L_2) Regularization, rather than using a single surrogate outcome. For the replacement algorithm's target allocation ratio, we modify Zhang & Rosenberger's (2006) continuous optimal allocation ratio for minimizing responses to a form more appropriate for repeated measures data. This modification involves a weighting system for the predicted final outcomes that is weighted less when there is less confidence in the predictions. Various weighting schemes and penalties based off time of individual's prediction (relative to their final outcome), current sample size, and missing data are considered. Preliminary results are presented comparing the design's performance metrics (under different conditions) against and an analogous RA design without the replacement algorithm.

Title: *Prognostic Modeling of Recovery Following Bone Marrow Transplantation*
Presenter: Brielle Forsthoffer
Advisor: Dr. Roy T. Sabo
Abstract:

A key concern in stem cell transplantation (SCT) is the determination of transplant prognosis. Current methods to classify patients according to lymphoid recovery depends on arbitrary criteria being set prior to modeling absolute lymphocyte counts (ALCs, μL^{-1}) over time. Because lymphoid recovery has been displayed as distinct courses individuals might take following myeloablative stem cell transplantation, we aim to utilize machine learning algorithms, not only to help make objective classifications, but to also prospectively predict the course an individual might take following SCT. Our approach is to use an adaptation of the group-based trajectory model (GBTM) using cubic B-splines to determine the clinical trajectories of any SCT recipient based off ALCs. First, diagnostics are used to determine the appropriate number of groups and a suitable data cut-off for existing patients. Then, we use these classification models to prospectively predict a new patient's trajectory each time the patient is measured. The unambiguous classifications made based off of probabilities of group membership can be used to help clinicians make decisions on whether to intervene on the patient's behalf and alter treatment. Various scenarios will be evaluated using simulation studies and compared in terms of group-specific misclassification rates and modal average posterior probabilities (APPs).

Title: *preciseTAD: A machine learning framework for precise TAD boundary prediction at base level resolution*
Presenter: Spiro Stilianoudakis
Advisor: Dr. Mikhail Dozmorov
Abstract:

Background: Chromosome conformation capture combined with high-throughput sequencing experiments (Hi-C) have revealed that chromatin undergoes layers of compaction through DNA looping and folding, forming dynamic 3D structures. Among these are Topologically Associated Domains (TADs), which are known to play critical roles in cell dynamics like gene regulation and cell differentiation. Precise TAD mapping remains difficult, as it is strongly reliant on Hi-C data resolution. Obtaining genome-wide chromatin interactions at high-resolution is costly resulting in variability in true TAD boundary location by conventional TAD calling algorithms.

Methods: To aid in the precise identification of TAD boundaries we developed preciseTAD, a random forest based framework that leverages the spatial relationship of high resolution ChIP-seq defined genomic elements, coupled with density-based clustering and scalable partitioning techniques. We benchmarked our method against a popular TAD-caller and a novel chromatin loop prediction algorithm on multiple cell lines.

Results: We show that our framework performs well when predicting both TAD and chromatin loop boundaries. Compared to established techniques, boundaries predicted by preciseTAD are more enriched for CTCF, RAD21, SMC3, and ZNF143 sites. Likewise, preciseTAD boundaries were more conserved across cell lines, highlighting their biological significance.

Conclusion: Our results outline strategies for predictive modeling of 3D genomic domains using 1D genome annotation data. The precise identification of TAD boundaries will improve our understanding of how epigenetics shapes the 3D structure of the genome.

Title: *R Package Development for Grouped Weighted Quantile Sum Regression Implemented in the Frequentist and Bayesian Frameworks*
Presenter: Matt Carli
Advisor: David Wheeler
Abstract:

Weighted Quantile Sum (WQS) regression was developed for the analysis of highly correlated, high dimensional chemical mixture data. Chemicals thought to be associated with a health outcome are identified through non-zero weights, with studies showing it to be more accurate in identifying such chemicals than traditional regression and regularization methods such as LASSO, adaptive LASSO, and elastic net. Grouped Weighted Quantile Sum (GWQS) regression extends this method to allow one to place chemicals of interest into groups such that different magnitudes and direction of associations can be determined for each pre-defined group. Our R package *groupWQS*, available on CRAN, implements GWQS in a frequentist framework. Parameters for the groups and their constituent weights are estimated through nonlinear optimization, where the weights are the weighted average of a user-defined number of bootstrap test statistics. Currently the package supports binary and continuous outcome variables. Our second R package accepted by CRAN, *bayesGWQS*, implements GWQS in a Bayesian framework. While it currently only supports binary outcomes, it has the additional functionality of imputing missing values due to chemical concentrations that are below the limit of detection, a common source of missing data in these analyses. The flexibility of Bayesian models allow for the estimation of group parameters, their weights, and any missing data in the same MCMC chain.

Title: *Replication Crisis: Defining Replication*
Presenter: Alicia Richards
Advisor: Dr. Robert Perera
Abstract:

Introduction: In 2015 a study titled, "Estimating the Reproducibility of Psychological Science," replicated 100 psychology studies and found shockingly low reproducibility rates. Since the article was published, multiple researchers have suggested factors that influenced the low rates including publication bias, sample size, and underpowered studies. However, the weakness focused on in this study is how replication was defined and how different definitions may impact replication rates. Therefore, the purpose of this study was to examine the various definitions of replication and the conclusions the different definitions draw.

Methods: Using the reproducibility data, p-values, Bayes Factors, and mitigated Bayes factors were calculated and assessed at various thresholds to determine replication rates. Meta-analyses were also conducted to ascertain replication rates. Additionally, simulation studies were conducted to examine the amount of studies expected to replicate. All the replication rates and methods were then compared and reviewed.

Results: Using various alpha levels (0.05, 0.01, 0.005, 0.001), the percentage of the original studies with significant p-values compared to the replication studies ranged from 33%-97% versus 20-36%. Of the original significant studies only 36-42% replicated. Similar to p-values, Bayes factors only replicated 24-35% and mitigated Bayes factors replicated from 26-39%. Contrasting, meta-analyses presented higher replication rates ranging from 36%-68%.

Discussion: Due to the low replication rates and weaknesses of the various measures used to define replication we found that a better statistical definition of replication is needed. Thus, in the future we plan to develop a replication method that produces higher replication rates while decreasing limitations.

Title: *Distributed Skewed and Heavy-Tailed Multivariate Longitudinal Regression Models*
Presenter: Reuben Retnam
Advisor: Dr. Dipankar Bandyopadhyay
Abstract:

Estimation of models for three-way data in the matrix-variate setting presents a favorable alternative to traditional methods such as stacking. However, issues present in the analysis of observational longitudinal data such as irregular observation times and a lack of a similar number of observations per subject often hold back direct application of off-the-shelf matrix-variate regression models. In our work, we first develop regression models for longitudinal data that utilize the matrix-variate skew-t distribution, while addressing the aforementioned challenges. The use of the Matrix-Variate Skew-t (MVSt) distribution in these models allows researchers to model skewed and heavy-tailed data in this matrix-variate setting, as opposed to traditional matrix-normal models.

We then scale these models to tackle large amounts of data mined from electronic health records. This scaling is achieved via simultaneous implementation of techniques such as divide-and-conquer and predictive acceleration in the stages of a conditional expectation-maximization algorithm.

Simulation studies demonstrating the efficacy of runtime-improvements on synthetic data and applications to a large observational epidemiologic dataset tracking periodontal disease in a diverse population will be demonstrated.

Title: *Detecting Wake After Sleep Onset using Actigraphy Data*
Presenter: Xinxin Sun
Advisor: Dr. Shanshan Chen
Abstract:

Wake after sleep onset (WASO) is a metric defined as the total time spent awake from sleep onset to final awakening. Increasing WASO is associated with lower sleep quality, aging, and other health outcomes. The gold standard for measuring WASO is polysomnography (PSG) in sleep laboratories, which limits its use for in-home, longitudinal monitoring. Actigraphy is a potential alternative allowing for longitudinal monitoring, however, its capability to detect WASO has yet to be explored. In this work, we extend our previously - developed hierarchical framework for sleep-cycle detection using actigraphy and investigate whether actigraphy alone can be used to detect WASO. Using the PSG labels from the MESA dataset and synchronized actigraphy data, we built several supervised and unsupervised classifiers , and evaluated the performance of these classifiers using common machine-learning metrics and Bland-Altman analysis. Although all models demonstrated improved accuracy and C-index over the null model, the proportional biases between the true WASO and the detected WASO were large, warranting post-calibration procedures. Our study suggests using the actigraphy sensor alone for WASO detection remains a challenge. Additional sensor or subject-specific information, or carefully-calibrated models are required to address it.

Title: *Estimation method for the semiparametric accelerated mixture cure model with informative cluster size*
Presenter: Jonathan W. Yu
Advisor: Dr. Dipankar Bandyopadhyay
Abstract:

In clustered data such as the United Network of Organ Sharing (UNOS) database, the center size can affect patient survival time after transplant with its access to medical resources. Improper statistical procedures to handle informative cluster size can lead to biased results and misleading inferences. While the accelerated failure time (AFT) mixture cure model and the Cox proportional hazards (PH) mixture cure model are two classic models to analyze clustered survival data, the AFT has attracted less attention than its semiparametric counterpart due to the complexity of the estimation method. However, its direct physical interpretation and developments to the rank-based generalized estimating equations (GEE) provides an incentive to use for censored failure time data. We propose an estimation method for the semiparametric AFT mixture cure model that employs a faster expectation-maximization (EM) algorithm, the SQUAREM or DAAREM, that can accelerate any fixed-point and smooth mapping with linear convergence rate and an induced smoothing inverse cluster size reweighting procedure to handle the informative cluster size. To evaluate the performance of the proposed method, we conducted a simulation study. The results of the simulation study demonstrate that the proposed method performs better than the existing estimation method. We apply the proposed method to UNOS data of failure times from kidney transplant patients to demonstrate that this approach has better numerical performance than existing methods in literature.

Title: *Providing a better tool for modeling risk of elevated blood lead levels across the United States*
Presenter: Joseph Boyle
Advisor: Dr. David Wheeler
Abstract:

Increasing attention is being paid to the role that elevated blood lead levels (EBLLs) play in childhood health. However, research is limited on a national scale by an absence of comprehensive blood lead testing data, as many states do not publicly provide this data. In addition, data that are provided are not reported on a common spatial scale. The goal of this project was to provide a new tool for estimating risk of elevated blood lead levels for ZIP codes in the United States that improves on the performance of existing tools (e.g., the Vox lead risk score). We used several approaches to model the proportion of blood lead tests that were elevated including weighted quantile sum (WQS) regression, Bayesian index models, and random forests. We considered a variety of demographic and socio-economic status (SES) variables and estimated a ZIP code level deprivation index with the WQS and Bayesian index models. With the WQS model approach, we considered multiple spatial scales (e.g., ZIP code, census tract, county) of test data as well as different number of variables in the index. Models were compared based on predictive performance on an independent testing set of ZIP code EBLL data via correlation between predicted and observed EBLL proportion and residual comparison. We applied the best-fitting model to all US ZIP codes and made available a risk map via a web app. Using this tool, those living in the US may more accurately assess their lead risk, and public health resources can be directed to posited "hot spots" of risk that are not yet publicly identified.

Title: *Power to Detect Main and Interaction effects of Two or More Treatment levels on a Binary Outcome in an Unbalanced Cluster Randomized Trial*
Presenter: Dongho Shin
Advisor: Dr. Yongyun Shin
Abstract:

This research aims to develop methods for accurate computation of power to detect the treatment effect on a binary outcome in a generalized linear mixed model (GLMM). Existing approaches linearize the binary outcome as a continuous one by the Taylor series expansion, also known as the marginal or penalized quasi-likelihood methods, and computes power based on the continuous outcome in a linear mixed model (LMM). Although these methods provide flexible means to compute the power based on widely known power tools in LMMs, they use ad hoc translation of the known or estimated parameters of the GLMM into those of the LMM and estimate the power based on the translated parameters. The produced power may not be accurate because it is based on the relationship between the two sets of the parameters of the GLMM and LMM that are not well known. For example, the translation between the between-cluster variances or the intra-cluster correlation coefficients of the two models has not yet been well studied. Therefore, in this project, we reveal how to characterize the parameters of the GLMM in terms of the parameters of the LMM by an extensive simulation study. Based on the found relationship, we plan to develop methods to compute power accurately by a principled approach. We will compare our approach against existing methods in terms of bias, type I error, and coverage probability as well as power to detect the treatment effect.

Title: *Sing-Index Cure Model*
Presenter: Jing Zhang
Advisor: Dr. Dipankar Bandyopadhyay
Abstract:

We work on cure models for time-to-event data. The survival function of a cure model includes two components: incidence and latency. The "incidence" represents the probability of being uncured and the "latency" refers to the survival function of the uncured observations. In literature, the incidence follows the logistic model while the latency follows the Cox proportional hazards model. EM algorithm is used to estimate the incidence and latency. Later, to cope with the "curse-of-dimensionality" problem, the incidence part is modeled by single-index model, the new model is referred as "single index/Cox mixture model". In our project, we modify this model by using single-index model for both the incidence part and the latency part. And EM algorithm is used to estimate these two parts. For the application of the models, we compare these three models using the Qatari Lung Cancer data set. The estimated uncured probability and the AIC/BIC for each model are calculated and compared.

Title: *Influence of clinical characteristics and patient demographics on aggressive end of life care*
Presenter: Dustin Bastaich
Advisor: Dr. Donna McClish
Abstract:

When patients are nearing the end of their life, the goal of treatment changes from extending life to providing comfort. Aggressive treatment can cause pain, discomfort, and negative effects to mental health without providing the benefit of meaningfully extending the patients time to live. The goal of this study is to examine factors that may be associated with aggressive treatment in end of life care.

This was a population based retrospective cohort study looking at all cancer patients in Virginia who died between 2012-2015. We examine aggressive treatment in the last 30 days of life, including: 2 or more hospitalizations, more than 14 days in the hospital. being in an Intensive Care Unit, receiving invasive procedures, or dying in hospital. The association of clinical characteristics (multiple cancers, hematologic diagnosis, previous hospitalizations, poor prognosis diagnosis, time from diagnosis to death, year of death) and patient demographics (age, race, gender, SES, rural residence) with each aggressive treatment in the last 30 days of life was examined through logistic regression models. Interactions between age group (>=65, <65) and all other factors were added to examine if factors differ based on patients meeting the age requirement for Medicare.

Patients receiving aggressive treatment tended to be male, black, low socioeconomic status, and have more hospitalizations in the past 2 years, a hematologic diagnosis, and a single cancer diagnosis. For many of the models created, patients being at least 65 years old was an effect modifier on multiple clinical characteristics and patient demographics.

Title: *Assessing the impact of censoring on the survival prognostic accuracy measure*
Presenter: Shannon Moldowan
Advisor: Dr. Le Kang
Abstract:

The Harrell's C-index is widely used to quantify the prognostic accuracy of biological markers or risk scoring algorithms given censored survival outcomes. However, it has been recognized that censoring may introduce biases to the estimation of such accuracy measure, resulting in misleading and incorrect inference on the prognostic abilities of biomarkers. In this project, we explored the simulations and Monte Carlo methods to generate censored survival outcomes together with the prognostic biomarkers given specified C-index, and we assessed the impact of censoring on the estimation of Harrell's C-index based on simulation studies. We demonstrated that the common strategy of comparing multiple biomarkers in terms of C-indices could be problematic through some numerical examples, e.g., in a time-varying covariate effects model.

Title: *A comparison of sample size calculation methodologies for Phase II Clinical Trials*
Presenter: Chen Wang
Advisor: Dr. Roy T. Sabo
Abstract:

Phase II trials, which are based on a moderate number of subjects to allow the research question to be addressed in a relatively short time, often allow a preliminary review of a new treatment before embarking on a more extensive randomized control trial. One common problem in phase II trials is that too many subjects are needed so that the results cannot be provided quickly. Many approaches have been used previously to address this problem, most notably Bayesian methods; however, a few drawbacks to these approaches include that (1) the trade-off between power, type I error and sample size remains unclear, (2) the sample estimates may be too variable early in the study, and (3) they may not terminate early enough. In this research, we compare Bayesian methods to those using the decreasingly informative prior (DIP) model for one-sample binary outcomes. Simulation studies are performed to measure the power, type I error, the end-of-trial sample size (standard deviation), and the proportion of completed trials for which the experiment treatment is promising.

Title: *An adaptive method for covariate balancing in block randomized clinical trials*
Presenter: Rebecca Rasnick
Advisors: Dr. Adam Sima and 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 he 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 trail design that limit randomization bias include stratified block design and propensity scores. To date, no adaptive methodology has been introduced to limit the randomization bias in block randomization trials.

Methods: The binomial distribution was used to calculate the expected final number of participants with a certain binary 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 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 future imbalance.

Results: Our proof-of concept design was able to maintain Type 1 error and power while decreasing the variability of the expected imbalance throughout each trial when the confounder had no effect on the trial.

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: *Robust quantile estimation of the Brief Test of Adult Cognition by Telephone (BTACT) battery*
Presenter: Serenity Budd
Advisor: Dr. Adam Sima
Abstract:

The Brief Test of Adult Cognition by Telephone (BTACT) is a cognitive battery featuring six tests used to assess an individual's cognitive functioning. The raw scores of these tests have previously been shown to be associated with an individual's age, sex, and education level. In order to use the BTACT battery to diagnose mild cognitive impairment, an individual's scores need to be compared to those from a healthy population after accounting for age, sex, and education level. Standard methodology estimates the distribution of the tests using a strict distributional assumption with categorized age levels. This project assess the normative distribution of the BTACT tests using more robust methodology. We utilized the Midlife in the United States data to build and independently validate our findings. A series of quantile regression models was fit to each BTACT test to find the predicted quantile level for a given participant's actual test score. The quantile regression models were evaluated by calculating the percentage of individuals with predicted quantiles less than 0.10, 0.25, 0.50, and 0.75, separately for independent training and test sets. This method was compared to a traditional z-score method. We demonstrated that the quantile regression method has characteristics more similar to those expected for the range of quantiles explored.

Title: *Survival with Total Artificial Heart (TAH)*
Presenter: Atika Farzana Urmi
Advisor: Dr. Leroy Thacker
Abstract:

The Mechanical Circulatory Support (MCS) devices can help patients with advanced or end-stage heart failure to restore blood flow, increase survival and improve quality of life. The two most commonly used MCS devices are the left ventricular assist devices (LVADs) and the Total Artificial Heart (TAH). The devices can serve as a bridge to transplant, destination therapy or bridge to recovery depending on the patient receiving it. However, with an increasing global burden of cardiovascular disease and congestive heart failure, the number of patients with end-stage heart failure awaiting heart transplantation now far exceeds the number of available hearts. As a result, the use of mechanical circulatory support is growing exponentially. Unlike LVAD, which is used for left ventricular failure, the TAH is used in patients with end-stage heart failure affecting both sides of the heart (biventricular failure).The current study was a single centered (VCU medical Center), retrospective study that includes 116 TAH primary implants .Primary outcome of the study was 6 month and 1 year overall Survival (Alive on TAH or Transplant).Secondary outcome includes 1, 3 and 5 year post transplant survival and overall survival . Predictors such as patient's age, ischemic heart disease, renal failure, serum bilirubin, center experience , intermacs profile etc. are used to model risk for survival to transplant and overall survival. The objective of the study was to examine the variables that influence TAH failure before transplant and effect of time dependent covariates (e.g. age) on cardiac patients' survival.