Biostatistics Student Research Symposium (BSRS) 2021

Abstracts

VCU Biostatistics
School of Medicine

Title: *A 'divide-and-conquer' EM Algorithm for large non-Gaussian Longitudinal Data with Irregular Follow-Ups*
Presenter: Reuben Retnam
Advisor: Dr. Dipankar Bandyopadhyay
Abstract:

Features of non-Gaussianity, manifested via skewness and heavy tails, are ubiquitous in databases generated from large scale observational studies. Yet they continue to be routinely analyzed via linear/non-linear mixed effects models under standard Gaussian assumptions for the random terms. In periodontal disease data, these issues are applicable to the modeling of clinical attachment level and pocket depth. These problems are maintained, if not exacerbated, in the longitudinal data framework.

In this research, we define and elucidate an extension of the skew-t linear mixed model suitable for a big data setting. This extensibility is achieved via the implementation of divide-and-conquer techniques that utilize the distributed expectation-maximization algorithm. Specifically, the E-step of the EM algorithm is run in parallel on multiple worker processes, while manager processes perform the M-step with an updated fraction of the results from the local expectation steps. We prove convergence properties of this algorithm and show examples of its performance compared to traditional modelling methods on real and simulated data.

Title: *Novel feature evaluation in ultra-high dimensional TCGA Head and Neck Cancer data with right-censored endpoints*
Presenter: Atika Farzana Urmi
Advisor: Dr. Dipankar Bandyopadhyay and Dr. Chenlu Ke
Abstract:

With recent explosion of ultra-high dimensional data of unprecedented size and complexity, feature screening and feature selection are playing an increasingly important role in scientific studies. In many large-scale biomedical studies generating gene expressions of very high dimensions (such as cancer), important "time-to-event" primary outcomes are subjected to right-censoring, creating impediments to efficient and robust gene selection while employing available methods. In this project, we develop a model-free gene screening procedure by modifying the expected conditional characteristic function-based independence criterion (ECCFIC) of Ke et al. (2019) to measure the dependence between each gene and survival outcome. The screening stage eliminates most irrelevant genes and fast reduces the ultra-high dimensionality to a moderate scale. Subsequently, classical penalized approaches are further applied in the post-screening stage for more precise gene selection and predictive modeling. The performance of our two-stage method was compared with other existing screening procedures based on the ROC plots and AUC.

Title: *The Replication Crisis: Reassessing, Rebuilding, and Redefining Replications*
Presenter: Alicia Richards
Advisor: Dr. Robert A. 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, researchers have suggested factors that may have influenced the low rates including publication bias, sample size, and underpowered studies among others. However, the current methods to define replication do not address these factors and only assess replication using binary definitions. Therefore, the purpose of this study is to assess the presence of these limiting factors using the reproducibility data in order to build realistic simulations to help redefine replication.

Methods: Using the Reproducibility Project's original and replicated study results and data, Monte Carlo simulations were performed to evaluate the impact power/sample size and publication bias play in the studies ability to replicate. Various levels of publication bias, power, and effect sizes (small, medium, large) were utilized to include a range of simulations.

Results: This analysis found that most of the original and replicated studies had insufficient power and a high presence of publication bias that is not desirable. Thus, using the various levels of power, bias, and variability we presented under idealistic simulation conditions were not realistic for this data since the levels of power and bias were substandard.

Discussion: In order to build replications that are realistic and idealistic a large range of power levels, effect sizes, and bias need to be accounted for. As we assess replication continuously, this study provides us with the guidance and justification for selecting simulation conditions.

Title: *Variability in Causal Effects of e-Assist on a Binary Outcome and One-Sided Noncompliance in a Multi-Site Randomized Trial*
Presenter: Xinxin Sun
Advisor: Dr. Yongyun Shin
Abstract:

Housed within a patient portal, e-Assist is a decision support program that assists patients with a physician recommendation for completing colorectal cancer screening (CRCS=1 if done, 0 otherwise). The patients are randomized to e-Assist or usual care within physicians. However, the physician-specific effects of e-Assist on CRCS vary randomly over physicians of varied composition, and compliance to treatment assignment is imperfect. Because a control patient is unable to access e-Assist and forced to receive usual care, the patient could be "complier" if she would have taken e-Assist or “never taker” otherwise under the alternative assignment to e-Assist; a patient assigned to e-Assist is an observed complier if the patient took the assignment or never-taker otherwise. Assuming compliance missing at random, we estimate a joint random coefficients model for CRCS and compliance by an accelerated combination of the EM algorithm and Newton Raphson via adaptive Gauss Hermite quadrature. Random coefficients are physician-specific never-taker and control-patient CRCS rates and complier average causal effects whose means, variances and correlations describe the population of compliers.

Keywords: compliance MAR; physician-specific causal effects, random coefficients, accelerated EM algorithm, adaptive Gauss Hermite quadrature, maximum likelihood.

Title: *Imputation of Below Detection Limit Missing Data in Chemical Mixture Analysis with Bayesian Group Index Regression*
Presenter: Matthew Carli
Advisor: Dr. David C. Wheeler
Abstract:

There is growing scientific interest in identifying the multitude of chemical exposures related to human diseases, including cancers, through mixture analysis. A method recently developed to conduct such joint analyses is Bayesian group index regression, which has been shown to have better power, specificity, and sensitivity than other methods such as the lasso and weighted quantile sum regression. A common problem in such analyses is the issue of below detection limit (BDL) missing data. We propose to treat both Bayesian group index regression effects and missing BDL observations as parameters in a MCMC algorithm we will refer to as Pseudo-Gibbs imputation. We compare this with other Bayesian imputation methods found in the literature (Multiple Imputation by Chained Equations and Sequential Full Bayes imputation), as well as with a non-Bayesian single imputation method. To evaluate our proposed method, we conduct simulation studies with varying percentages of BDL missingness and strength of association. Our results indicate that, compared to the other implemented methods, Pseudo-Gibbs imputation has superior power for exposure effects and sensitivity for identifying individual chemicals at high percentages of BDL missing data. Results also show that at low levels of BDL missingness, relatively simple single-imputation methods have comparable performance to more sophisticated and computationally intensive methods. In conclusion, Pseudo-Gibbs imputation addresses a commonly encountered problem in environmental epidemiology, providing practitioners the ability to jointly estimate the effects of multiple chemicals with high levels of BDL missingness.

Title: *Knot selection for low-rank kriging models of spatial risk in case-control studies*
Presenter: Joseph Boyle
Advisor: Dr. David C. Wheeler
Abstract:

Case-control studies have been used to map spatial distributions of disease and detect potential cancer clusters. Using individual point data allows precise detection of geographic areas of elevated risk in spatial regression models but begets a large computational burden. Low-rank kriging (LRK) is a computationally more efficient method than regular kriging and generalized additive models because it uses a reduced set of points known as knots for estimating spatial random effects. A common approach to select knots is a space-filling coverage algorithm, which minimizes a geometric criterion over a discrete grid. However, the efficacy of this approach is unknown in case-control studies, where the spatial distribution of cases and controls may differ. We present a simulation study that compares the commonly-chosen knot selection method with two others that consider case-control status: the same method using only cases, and the Teitz-Bart location-allocation algorithm. The latter two methods seek to represent the spatial distribution of cases and may enable greater power to detect areas of disease risk. We simulate population distributions in a study region and generate zones of elevated risk and then compare the spatial sensitivity, specificity, and power of LRK models that employ these knot selection methods. We find that the knot selection methods that consider the case-control status of participants perform much better than the commonly-used method, more frequently identifying areas of elevated risk. Finally, we apply the Teitz-Bart method to a LRK model of New England Bladder Cancer Study participants to identify potential regions of elevated spatial risk.

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

Adaptive allocation adjusts the ratio of participants assigned to the better performing treatment arm in a clinical trial by using outcome data from earlier accrued participants. This offers an ethical advantage compared to randomized controlled trials with equal allocation by reducing participant allocation to potentially inferior treatment(s). However, if the time-to-response is too long, then adaptive allocation cannot effectively materialize this advantage. 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. However, it is not uncommon for studies of long duration to have repeated measurements of the primary outcome (in part) to monitor the progress of study participants. Thus, we propose a replacement algorithm that implements earlier repeated measures of the primary outcome to predict the final outcome via curvilinear growth models. 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 downweighting predicted outcomes to account for their uncertainty. Various weighting schemes and growth models are considered and compared under a variety of conditions including different effect sizes, accrual rates, and outcome change patterns. The proposed method is also compared against an analogous response adaptive design without the replacement algorithm.

Title: *Needs Assessment of Health Inequities in Eastern Henrico*
Presenter: Emily Walzl
Advisor: Dr. Tegwyn Brickhouse and Dr. Caroline Carrico
Abstract:

Identifying factors that contribute to use of community and health care services can better establish what community-based interventions would best improve overall health for parents and their children. A community survey was sent out to residents of zoned neighborhoods in the eastern region of Henrico County. Participants were asked about their and their children's health status, and utilization of health and community services. This survey also assessed factors that may contribute to their health and habits such as housing, transportation, and employment. The use of community services pre-Covid (prior to March 2020) were compared to the use of services at the time of the survey (July 2020 thru December 2020). Using a Mcnemar's Test, we compared the frequencies of respondents who stated they used or didn't use particular community services prior to Covid versus when the survey was taken. We found that for certain services there were significant differences in the distributions of service use prior to Covid and currently (p-value <.05). A Chi-square test was used to determine frequency differences in people's employment status and their use of services; a significant difference was found for participants' use of dental care services during the time the survey was taken (p-value <.05). A logistic regression model was created to predict the use of a community service; time of use (either currently or pre-Covid) was used for the exposure variable and demographic variables were used as covariates in the model. The goal is to use the results of this survey to better determine factors that allow families to access certain services in their neighborhood as well as determine what services the families deem valuable for their health.

Title: *Panel Size Estimation in Primary Care*

Presenter: Martin Lavallee
Advisor: Dr. Roy T. Sabo

Abstract:

Primary care is on the frontlines of the healthcare system, helping patients access medical care, provide consultation, and advocate public health initiatives. A well-functioning primary care arm seeks to maximize continuity and access to care, which can be accomplished through empanelment. Empanelment, the process of linking patients to providers, allows a practice to quantify panel sizes used to monitor the balance between patient demands and the capacity to provide care. Estimating panel size is not as simple as counting the number of patients seen by a provider because each patient is different in terms of complexity (i.e., presence of chronic diseases or other risk factors) and utilization (i.e., some patients frequently seek care while others only attend routine checkups). A weighted count of patients is a more equitable estimate of panel size where the weight is the relative burden posed by the patients based on features of utilization and complexity. The goal of this project is to explore and assess different unsupervised learning algorithms for estimating panel size that consider variation in patient healthcare features. We present two types of methods, one for utilization only data and a second for a mixture of complexity and utilization data. Within each method set we explore three types of approaches: k-means clustering, gaussian mixture model and a novel method called patient burden scoring. We consider how well these 6 methods do for estimating panel sizes using a real-world application from VCU Health systems data.

Title: *Exploring propensity score techniques in NHANES survey data*
Presenter: Rasnick
Advisor: Dr. Dipankar Bandyopadhyay
Abstract:

While large nationwide surveys, such as the NHNANES, can provide excellent insight into population-level estimates, the derived estimates of causal effects of variables (such as Race and periodontal disease status as in our case) on a desired response (such as oral cancer screening) can be biased, mostly in presence of confounders. It is well understood that propensity score (PS) techniques can be employed in these situations to mimic results as derived from a randomized controlled trial. However, while utilizing the popular inverse probability of treatment weighted (IPTW) technique to incorporate PS in a regression setup, it is not clear how to combine the IPT weights with the survey weights. In this project, we compare and contrast three popular techniques, viz., (a) the generalized boosted model (GBM), (b) the covariate balancing PS, and (c) the standard logistic regression, for computing the PS, and then follow the suggestions of Dr. Ridgeway (cite his paper here) regarding weighting to study the causal effects of race and PD status on oral cancer screening. In the process, covariate balancing was checked, and results were reported in terms of odds ratios, and associated 95% confidence intervals.

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

We compare standard Bayesian methods with an informative yet skeptical prior Bayesian method – the decreasingly informative prior (DIP) method – for early termination in one-group and two-group Phase II Clinical Trials. Comparisons between these two methods are made for binary and continuous outcomes. We want 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 simulate observed data and search through all possible values of total sample sizes (N) and decision criteria (p_f, p_s) to find the smallest total sample size (N) under the admissible power and type I error. Simulation studies in the one-group trial show: the DIP approach can have similar power and type I error with fewer patients, though it is achieved with better control of type I error than the standard Bayesian methods. We will also discuss the simulation results for the two-group trial. The results help to pre-determine termination rules for each observed total sample size.

Title: *Penalization Variable selection Methods for Competing Risks: An Application to UNOS Kidney Transplant Data*
Presenter: Edem Defor
Advisor: Dr. Dipankar Bandyopadhyay
Abstract:

Transplant centers often encounter events that preclude the occurrence of the primary event of interest. Rather than censoring, this is a competing risk. Our research focuses on kidney transplants, where the recipient's death before transplant prevents the transplant from occurring. We describe variable selection and parameter estimation methods in the proportional subdistribution hazard (PSH) model proposed by Fu et al. (2017) for competing risks. The usefulness of the PSH model is also presented. Using the novel backward selection scan algorithm in the fastcmprsk package in R, we demonstrate an impressive computational efficiency over existing packages to select important variables. Data on kidney transplants between 1999-2020 from the United Network of Organ Sharing is utilized.

Title: *Evaluation of Multiple Imputation by Chained Regressions (MICE)*
Presenter: Dongho Shin
Advisor: Dr. Yongyun Shin
Abstract:

This research evaluates MICE for estimation of a multilevel or hierarchical linear model given partially observed outcome and covariates under the assumptions of data missing completely at random and missing at random. A theoretical consideration sheds light on the joint distribution of the partially observed variables to which MICE converges. Simulation studies will distinguish cases when MICE produces unbiased and efficient estimation from cases when MICE produces biased or inefficient estimates.

Title: *Selection Model for COVID-19 recovery and Informative Dropout Given Web Survey Data MNAR*
Presenter: Serenity Budd
Advisor: Dr. Yongyun Shin
Abstract:

A longitudinal survey was conducted for COVID-19 patients who suffered from smell loss with the aim of exploring recovery. The outcome of interest was whether the participant regained a normal sense of smell at the six-month follow-up. Sixty-five percent of participants did not respond and, thus, had unit nonresponses at the six-month follow-up survey. The goal of this project is to understand and model the joint distribution of binary recovery outcome and dropout status at the six-month follow-up given covariates at baseline. Following preliminary analysis, we reason that the outcome, the six-month recovery outcome, was missing not at random and if patients recovered their sense of smell, they would be more likely to drop out of the study. We analyze the selection model that decomposes the joint distribution into dropout conditional on recovery and marginal recovery, controlling for baseline covariates. Existing software, however, cannot estimate the selection model. We developed the Newton-Raphson algorithm and coded it in R to estimate the parameters of the selection model. We evaluate the method by simulation and present results from its application to the COVID-19 data.

Title*: Improving machine learning modeling and predictions of 3D domain boundaries*
Presenter: Khoa Huynh
Advisor: Dr. Mikhail Dozmorov
Abstract:

Chromosome conformation capture techniques such as Hi-C has shown that the genome of many species is organized and known as topologically associating domains (TADs). The preciseTAD algorithms utilized random forest model trained on high-resolution genome to identify position of boundary TADs at base-level resolution with probability close to one. However, it missed the less significant but prominent regions with high boundary likelihood. To improve precision of preciseTAD, we utilize local weighted (loess) regression to smooth the probability vector from preciseTAD algorithms. Local extrema will be classified by inflection point from second derivative of smoothing lines. Boundary of TADs consider as local maximum, and midpoint length of TADs consider as local minimum. Since TADs are symmetric at midpoint of the length of TADs, our algorithm recognizes TADs if difference of first position local maximum and position local minimum approximately equal to difference of position local minimum and position local maximum. We show that our algorithm can identifying TADs with either high probability or low probability to be boundary. Various scenarios of TADs will be evaluated in our algorithm. The output algorithm was able to give the position of TADs boundary in genome data as well as the size of TADs.

Title: *Using deep learning to classify polyp detection and polyp retrieval from clinical colonoscopy reports*
Presenter: Dustin Bastaich
Advisor: Dr. Bassam Dahman
Abstract:

Information on polyps can be found in the clinical notes of colonoscopy reports, but it can be a time-consuming process to manually find this information through chart review. Deep learning classification could be an efficient method of extracting polyp information from clinical reports without the need for researchers to manually review each report. This could lead to higher sample sizes used in studies as well as faster completion time.

The reports in this study came from Virginia Commonwealth University (VCU) electronic medical records, where surgical clinical notes for all colonoscopy procedures performed at VCU Health from 2010-2012 were included. After removing reports from the same patient in the same year and those that had missing polyp detection and retrieval outcomes, we were left with 3463 clinical reports for analysis.

Neural networks with a word embedding and LSTM layer were used to classify reports with polyp detection and polyp retrieval. The classification of the two outcomes was handled separately, and 8 separate structures were fit with varying complexity. Model performance was compared based on accuracy, sensitivity, specificity, recall and F-1 score calculated from test data. The networks fit made use of randomized initial weights, so sensitivity analysis using 30 runs of each model was performed to examine how much performance varied with different initial weights.

The highest performing models had 97.8% classification accuracy for polyp detection and 96.5% accuracy for polyp retrieval. Sensitivity analysis showed that the change in performance from different initial weights can lead to choosing different best performing structures.

Title: *Missing data interpolation in meta-analyses with disparate covariate information*
Presenter: Rasha Alsaadawi
Advisors: Dr. Ekaterina Smirnova (VCU) and Dr. Bryan Lau (JHU)
Abstract:

Meta-analysis is a popular approach for increasing the power by combining data from several independent studies. The NIH supported Environmental influences on Child Health Outcomes (ECHO) initiative is to-date the largest program launched to understand the factors that affect child development health and finding ways to enhance it. One of the main goals of (ECHO) program is to conduct large population-scale analyses by integrating data from multiple cohorts. However, linking and harmonizing data from heterogenous study populations is a methodologically challenging. For a given analysis, some studies may not collect information on one or two major covariates of interest, which leads to cohort-level missing data. Current statistical methods designed for cohort-level missing data interpolation assume that the covariates of interest have similar distributions across all studies, which may often be violated in real data. Thus, there is an urgent need to (1) assess whether multiple cohorts have similar distributions on covariates of interest for the meta-analysis study; and (2) identify specific cohorts that most similar to the cohort that is missing an important covariate. One approach is to use the covariates of interest to predict cohort membership by multivariate random forest models, and then utilize the average similarity measure between terminal nodes on the random forest trees to cluster observations with similar covariates. Thus, poor classification of cohort membership would suggest inability to predict cohort membership and thus similarity of joint covariates distribution. In this work, we conduct a simulation study to test accuracy of this approach and its effect on missing data imputation.