SORA-TABA-DLSPH RESEARCH DAY

Virtual Poster Presentations

University of Toronto
Toronto, Ontario
May 12th, 2022

Event Sponsors:

Dalla Lana School of Public Health

SORA

The Applied Biostatistics Association

ASA American Statistical Association

Société statistique du Canada / Statistical Society of Canada

York University / Université York

SickKids Research Institute

McDougall Scientific — Insights You Can Trust

Contents

Welcome

On behalf of the Organizing Committee it is my distinct pleasure to welcome you to the Annual SORA-TABA Workshop and DLSPH Biostatistics Research Day.

The event is intended to bring together the regional and local statistical communities who are interested in biostatistics and other applied areas of statistics and represents a joint effort between the DLSPH (Dalla Lana School of Public Health), SORA (Southern Ontario Regional Association of the Statistical Society of Canada & Southern Ontario Chapter of the American Statistical Association), and TABA (The Applied Biostatistics Association). This year due to the continued COVID restrictions both the workshop and the poster presentations will occur virtually. The DLSPH Biostatistics Research Day Poster Presentations will take place over two separate sessions on May 12th. This will be followed by the SORA-TABA Workshop on Regression Diagnostics, featuring Professor John Fox from McMaster University, which will take place on May 19th and May 20th, 2022 from 11:00 to 17:00 EDT.

I would like to thank our sponsors for their continued support of this annual event. A special thank you is extended to Professor Emeritus Paul Corey; whose generous donation enables us to award monetary prizes to three students judged to have the best posters, and to the judges involved in the assessment of poster presentations. Finally, I would like to thank you for your participation and support of our program, especially given the changes necessitated by the COVID pandemic.

We hope you can join us.

Tony Panzarella (Chair of the Organizing Committee)

Special thanks to the representatives from SORA-TABA-DLSPH:

-Wendy Lou, University of Toronto (DLSPH)

-Lorinda Simms, Regulator, Biostatistics & Data Management at Partner Therapeutics (TABA)

-Tony Panzarella, University of Toronto (SORA)

Dedication

This event is dedicated to the memory of Janet McDougall, who died on June 13, 2021. Janet was the founder of McDougall Scientific Ltd., a long-time sponsor and supporter of the SORA-TABA Workshop & DLSPH Biostatistics Research Day.

As her obituary aptly stated *"She believed in giving back, paying forward, and in sharing whatever she had with heartfelt generosity".*

She will be greatly missed.

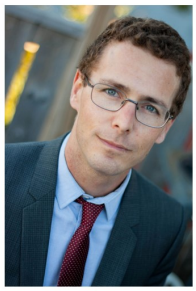
Organizing Committee

Tony Panzarella
(DLSPH,
University of Toronto)

Hugh McCague
(York University)

Teresa To
(The Hospital for Sick
Children)

Lisa Avery
(Princess Margaret
Hospital & SORA)



Ryan Rosner
(DLSPH,
University of Toronto)

Ruth Croxford
(ICES)

Kevin McGregor
(York University)

Myrtha Reyna
(The Hospital for Sick
Children)

Rose Garrett, PhD Candidate
(DLSPH,
University of Toronto)

Jiayin Chen, MSc Candidate
(DLSPH,
University of Toronto)

Keynote Speaker - Professor Emeritus John Fox

John Fox is Professor Emeritus of Sociology at McMaster University in Hamilton, Ontario, Canada, where he was previously the Senator William McMaster Professor of Social Statistics.

Professor Fox received a PhD in Sociology from the University of Michigan in 1972. He is the author of many articles and books on statistics, including, recently, Applied Regression Analysis and Generalized Linear Models, Third Edition (Sage, 2016), Using the R Commander: A Point-and-Click Interface for R (Chapman & Hall, 2018), Regression Diagnostics, Second Edition (Sage, 2019), A Mathematical Primer for Social Statistics, Second Edition (Sage, 2021), and, with Sanford Weisberg, An R Companion to Applied Regression, Third Edition (Sage, 2019).

He continues to work on the development of statistical methods and their implementation in software. Professor Fox is an elected member of the R Foundation for Statistical Computing and an associate editor of the Journal of Statistical Software.

Judges

- **Rosane Nisenbaum**
  Applied Health Research Centre & Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto

- **Kuan Liu**
  Institute of Health Policy and Management & Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto

- **Lisa Avery**
  Department of Biostatistics, Princess Margaret Cancer Centre & Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto

- **Alistair Johnson**

  Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto & Hospital for Sick Children

- **Ruth Croxford**

  ICES

- **Jenna Sykes**

  Department of Respirology, St. Michael's Hospital & Dalla Lana School of Public Health, University of Toronto

- **Aya Mitani**

  Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto

- **Sandra Gardner**
  Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto & Kunin-Lunenfeld Centre for Applied Research and Evaluation, Rotman Research Institute, Baycrest Health Sciences

- **Xuan Li**
  Department of Biostatistics, Princess Margaret Cancer Centre, University Health Net-work

- **Charles Keown-Stoneman**

  Applied Health Research Centre & Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto

- **Divya Sharma**

  Toronto General Hospital, University Health Network & Dalla Lana School of Public Health, University of Toronto

Presentation Schedule Session 1

Thursday, May 12th - 10:00 - 11:30am EDT

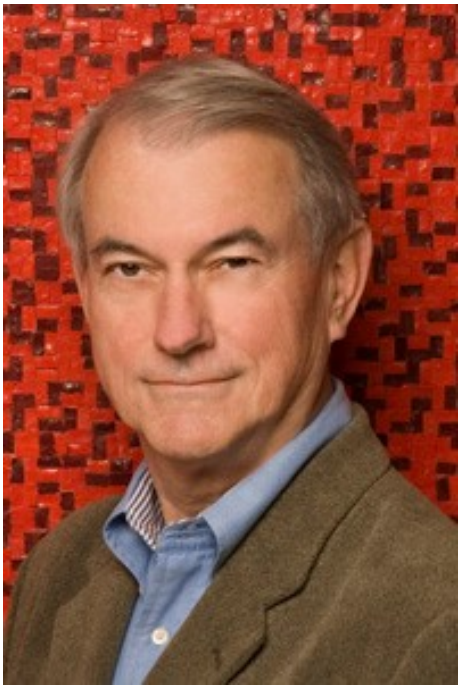
No	Title	Presenter	Time
1	Penalized maximum likelihood inference under the Mixture Cure model	Changchang Xu	10:05-10:10
2	Multi-omics integration via similarity network fusion to detect subtypes of cognitive aging	Mu Yang	10:10-10:15
3	Pulmonary and nutritional outcomes in CFSPID cohort	Maria Sbirnac	10:15-10:20
4	Multivariate Trajectory Models in Phenotyping Osteoarthritis	Walid Maraqa	10:20-10:25
5	A Novel Umbrella Trial Design for Dementia Prevention	Zijin Liu	10:25-10:30
6	Communicating COVID-19 risk with the public	Amirpooya Sadeghi	10:30-10:35
7	Defining Lifestyle Patterns for Pre-school-aged Children	Xiaotong (Emily) Liu	10:35-10:40
8	Machine learning approaches for classifying credit card clients based on default risk	Mei Han	10:40-10:45
9	Relationships between Major Health Behaviors and Sleep Problems: Results from 2017-2018 Canadian Community Health Survey	Mo Zhou	10:45-10:50
10	Sex-stratified vs sex-combined analysis in the presence of genetic effect heterogeneity	Boxi Lin	10:50-10:55
11	Dealing with Partially Observed Confounders and Feature Selections in Propensity Score Analysis: A Comparison of Different Approaches with Applications in Non-alcoholic Fatty Liver Disease	Yun Zhu	10:55-11:00
12	Childhood-onset Systemic Lupus Erythematosus: Defining Long-term Outcomes in Ontario	Ha-Seul Jeoung	11:00-11:05
13	Feature selection with ElasticNet to assess relationship between sleep and depression treatment	Michelle Wu	11:05-11:10
14	Nursing Involvement in AI Research	Mark Germano	11:10-11:15
15	Multiple Imputation Methods for Multilevel Ordinal Outcomes	Mei Dong	11:15-11:20

Presentation Schedule Session 2

Thursday, May 12th - 14:00 - 16:00 EDT

No	Title	Presenter	Time
1	Identifying characteristics and strategies for long-term success in the management of pediatric systemic lupus erythematosus	Fatema Tuj Johara	14:05-14:10
2	Integrating Genetic & Clinical Risk Factors on UKBiobank data in predicting NAFLD Risk using Machine Learning Algorithm	Jiayin Chen	14:10-14:15
3	A Comparative Study of Robust Estimation Methods for Cox Regression Model	Shiyao Ying	14:15-14:20
4	Identifying informative subsample selection strategies for shotgun metagenomics analysis in microbial studies	Daniel Felipe Segura Hincapie	14:20-14:25
5	Investigating the association between military length of service and rate of emergency department visits following release	James Saunders	14:25-14:30
6	Workflow for Identifying Protein of Interest	Jiayue (Irene) Feng	14:30-14:35
7	A Gaussian mixture model-based variational graph autoencoder algorithm for clustering single-cell RNA-seq data	Eric Lin	14:35-14:40
8	Assessing longitudinal trends in online brain health assessment scores for older adults living in community settings	Luis Ledesma	14:40-14:45
9	Evaluate the effect of data imputation on clustering analysis of single-cell RNA sequencing data	Boyuan Liu	14:45-14:50
10	Using classification and regression trees to model missingness in youth BMI, height, and body mass	Amanda Doggett	14:50-14:55
11	Identifying Genetic Determinants of Absence Seizure Incidence in Juvenile Myoclonic Epilepsy	Eric Sanders	14:55-15:00
12	Assessing the Impacts of Serial Correlation on the Performance of Structural Change Detections Using Simulation Method	Zixin Zhou	15:00-15:05
13	Are Vision and Gait Associated with Fear of Falling Among Patients with Neurodegenerative Disease?	Lee Radigan	15:05-15:10
14	Bayesian Cox Proportional Regression with Partial Likelihood	Ziang Zhang	15:10-15:15
15	COVID-19 public health measure adherence amongst parents and children: investigating the importance of sociodemographic factors during lockdowns and reopenings	Kevin Dang	15:15-15:20
16	A machine learning approach to differentiate between COVID-19 and influenza infection using synthetic infection and immune response data	Suzan Farhang-Sardroodi	15:20-15:25
17	Machine Learning for Radiologic Series Categorizations of CT Scan	Suvd Zulbayar	15:25-15:30
18	Maybe we should stay home: Incorporating behavioural change into epidemic models	Madeline A. Ward	15:30-15:35
19	A Comparison of Methods for Bayesian Inference in Clinical Trials	Ziming (Jocelyn) Chen	15:35-15:40
20	Big Data Clustering of SARS-CoV-2 Spike Glycoprotein Sequences	Vadim Tyuryaev	15:40-15:45
21	Deidentification of Free-text Clinical Notes	Peiqing Yu	15:45-15:50

Professor Emeritus Paul Corey



Professor Emeritus Paul Corey began his career at what is now DLSPH in 1968, teaching Biostatistics to students in the clinical and health sciences, applied simulation methods, as well as online methods of teaching statistics. He won teaching awards and was beloved by his students and colleagues here at DLSPH. He was known for providing guidance to his students and mentees, and for being generous to them with his time.

Professor Emeritus Paul Corey received his BSc in 1962 and his MA in Human Genetics in 1965, both from U of T, and completed a PhD in Biostatistics at Johns Hopkins in 1974. His research was vast but focused primarily on analysis of environmental and occupational health and nutritional science. He was also a Professor in the Department of Statistical Sciences at the Faculty of Arts & Science. He officially retired in 2016 but can still be found teaching on campus.

Abstracts for Poster Presentations

Abstracts Session 1

Thursday, May 12th - 10:00 - 11:30am EDT, Chair: Tony Panzarella

Click here for Zoom link

Penalized maximum likelihood inference under the Mixture Cure model

Changchang Xu, Shelley B. Bull

Introduction & Objectives: When a study sample includes a large proportion of long-term survivors, mixture cure (MC) models that separately assess biomarker associations with long-term recurrence-free survival and time to disease recurrence are preferred to proportional-hazards models. Standard maximum likelihood (ML) may be biased in small or sparse samples (i.e. with few recurrences). We aim to improve parameter estimation and inference for MC under such scenario.
Methods: We extend Firth-type penalized likelihood (FT-PL) developed for bias reduction in the exponential family to the Weibull-logistic MC, using the Jeffreys invariant prior. Via simulation studies based on a motivating cohort study, we evaluate parameter estimates of the FT-PL method in comparison to ML, as well as type 1 error (T1E) and power obtained using likelihood ratio statistics.
Results & Conclusion: In samples with relatively few events, the Firth-type penalized likelihood estimates (FT-PLEs) have mean bias closer to the true value and smaller mean squared error than maximum likelihood estimates (MLEs), and can be obtained in samples where the MLEs are infinite. Under similar T1E rates, FT-PL consistently exhibits higher statistical power than ML in samples with relatively few events. In addition, we compare FT-PL with parameter estimates obtained via two other penalization methods (a log-F prior method and a modified Firth-type method) based on the same simulations. Finally, the practicality and strength of FT-PL for MC analysis is illustrated in a cohort study of breast cancer prognosis with long-term followup for recurrence-free survival.

Multi-omics integration via similarity network fusion to detect subtypes of aging

Mu Yang, Stuart Matan-Lithwick, Yanling Wang, Philip De Jager, David Bennett, Daniel Felsky

Introduction & Objectives: Molecular subtyping of post-mortem brain tissue in cognitive aging, particularly with Alzheimer's disease (AD), has typically focused on single data modalities, such as RNA sequencing (RNAseq), which provide incomplete neurobiological information. We applied similarity Network Fusion (SNF), a method capable of integrating high-dimensional multi-'omics data modalities simultaneously.
Methods: We analyzed human neocortical brain tissue samples characterized by five modalities: RNAseq, DNA methylation, histone acetylation, proteomics, and metabolomics. SNF followed by spectral clustering was used for subtype detection. Normalized Mutual Information (NMI) was calculated to determine the contribution of each modality and feature to the fused network. Subtypes were characterized by associations with 12 age-related neuropathologies and cognitive performance.
Results: Fusion of all five data modalities (n=111) yielded four molecular subtypes among which episodic memory performance proximal to death differed significantly (p=1.5x10-4). Histone acetylation and RNAseq were the most influential modalities; top contributing features were an acetylation peak in the promoter of CD200 and RNA abundance of PARP4. Secondary analysis fusing only RNAseq and histone acetylation (n=520) yielded five subtypes which were correlated with the fully integrated subtypes and associated with AD neuropathology and episodic and semantic memory. Sensitivity analyses found notable influences of sample size and subtype number.
Conclusion: We identified highly integrative molecular subtypes of cognitive aging using up to five multi-'omic modalities simultaneously. These subtypes recapitulate some features of previous unimodal subtyping work in AD, but also provide new molecular targets and shed light on the benefits and challenges of multi-omic integration and individual subtyping in this field.

Pulmonary and nutritional outcomes in CFSPID cohort

Maria Sbirnac, Annie Dupuis, Tanja Gonska

Introduction & Objectives:Cystic fibrosis (CF) is an autosomal recessive condition that affects around 1 in 3300 neonates in Canada. In recent years, CF Newborn screening (NBS) programs, involving an initial immuno-reactive trypsinogen (IRT) measurement, have become popular in regions with high CF incidence. Those NBS programs have helped identify a subgroup of infants who present an inconclusive CF diagnosis despite reporting an elevated IRT. Those CF screen positive inconclusive diagnosis (CFSPID) infants may later transition to a CF diagnosis based on elevated sweat chloride concentration or reinterpretation of genetic data. NBS IRT level or trypsin decline may function as predictive markers for CFSPID infants at risk of later transition to a CF diagnosis. However, quantifying change in trypsin levels over time is difficult due to heterogeneity of laboratory methodology. Objective: To develop a method of standardizing decline of trypsin over time with values collected via different laboratory methods.
Methods: Those standardized scores may be incorporated as predictors in Cox proportional hazard survival models to quantify the effect of trypsin on CF diagnosis time-to-event data.
Anticipated Results: We are expecting decline of trypsin over time to be significantly greater for pancreatic insufficient CF patients than for pancreatic sufficient CF or CFSPID patients. As such, decline of trypsin over time may be used to identify pancreatic sufficient patients at risk of developing pancreatic insufficiency later in life.
Conclusion: We are anticipating our work will bring greater value to patient data and serve as starting point for development of improved CF risk prediction models helping children affected by CFSPID.

Multivariate Trajectory Models in Phenotyping Osteoarthritis

Walid Maraqa, Osvaldo Espin-Garcia, S. Amanda Ali

Intro & Objectives: Osteoarthritis is a degenerative disease of the joints impacting their structure and function as well as the individual's mobility and pain. One of the shortcomings of current methods of phenotyping OA by both its structural (physical, physiological, anatomical) artifacts and symptoms is the lack of concordance between the two (Patients report pain symptoms without structural artifacts and vice-versa)
Methods: We use longitudinal k-means clustering to to outline discrete pathways for OA, pathways that take into account both symptoms and structural changes rather than assume uniformity of progression. This form of k-means clustering makes use of metrics and distance functions to quantify differences in states at different time-points to find the nearest cluster to a data-point.
Results: We have been able to produce clusters in the dataset from the Osteoarthritis Initiative based on structural and patient-recorded symptoms (joint space width and pain, respectively). The quality of the cluster was judged based on a measure of between-cluster variance to within-cluster variance.
Conclusion: Despite there being many other variables in the OAI database whose value has not been examined in longitudinal k-means clustering, our initial findings show promise in the adaptability and utility of this novel procedure.

A Novel Umbrella Trial Design for Dementia Prevention

Zijin Liu, Clement Ma, Tarek Rajji

Introduction & objective: Despite the increasing incidence of dementia worldwide, up to 40% of dementia cases can be prevented if we develop interventions targeting 12 modifiable risk factors for dementia. Existing dementia prevention trials typically use the parallel-group randomized design. Novel trial designs, such as biomarker-stratified umbrella designs, can improve trial efficiency. To our knowledge, umbrella designs have never been used for dementia prevention. We propose a novel umbrella design for dementia prevention.
Methods: We considered two risk factors, smoking and hypertension, as "biomarkers" to stratify participants into three subgroups: (1) smoking only; (2) hypertension only; and (3) smoking and hypertension. Within each subgroup, we randomized patients into two arms: Integrated Care Pathway (ICP) vs. Treatment as usual. This design can assess subgroup and overall effects simultaneously. We conducted a simulation study to assess the power of our umbrella design to detect subgroup and overall effects. We varied the total sample size, subgroup allocation ratio, effect sizes, and interaction effects between ICP treatments across 7,200 scenarios.
Results: A fairly small overall sample size (n=60) provides >80% power to detect intermediate effect sizes (β=0.5) overall. There is also acceptable power (54%-97%) to detect subgroup effects for larger effect sizes (β=0.8) and sample size (n=180). Interaction effects between ICP treatments can also influence the overall power.
Conclusion: Umbrella trials are feasible and more efficient compared to standard randomized trials for dementia prevention. However, investigators need to carefully consider the interaction effects and subgroup allocation scheme to achieve the trial objectives.

Communicating COVID-19 risk with the public

Amirpooya Sadeghi, Daniel Sanchez Morales

Introduction and objective: Today's state of infodemic along with the rapid growth and changes of COVID-19 imposed a significant challenge on communicating the data and risks regarding COVID-19. Along with this challenge comes an opportunity to learn and improve such communication. This research is an attempt to evaluate the current methods of communicating COVID-19 data and risk with the public and identify better alternatives to improve the public's understanding of COVID-19 risks and uncertainties.
Methods: We looked at the Public Health Ontario and the British Columbia Centre for Disease Control to evaluate current methods and tools used to communicate COVID-19 data. We then searched through the works of David Spiegelhalter and Emily Oster to understand how we can improve communication of risks and uncertainties regarding COVID-19.
Results: Both public health authorities heavily rely on line and bar graphs along with numerical and verbal reports. Visualization of COVID-19 data needs to be enhanced through different methods such as the utilization of human graphs to improve the public's understanding. Further, the COVID-19 risk calculator developed by Emily Oster can help with people's awareness of risks and uncertainties regarding their environment and interactions. Additionally, a more coherent and uncluttered User Interface (UI) can significantly improve people's understanding of COVID-19.
Conclusion: The challenges of risk and data communication regarding COVID-19 have become evident. It is of great importance to reflect upon what has been done and what improvements can be made to help the public's understanding of data, risks, and uncertainties. In this poster, we evaluate some current methods of communicating COVID-19 data and risk with the public and provide alternatives with the aim of improving the public's understanding.

Defining Lifestyle Patterns for Pre-school-aged Children

Xiaotong(Emily) Liu, Myrtha E. Reyna, Zihang Lu, Wendy Lou

Introduction & Objectives: Multi-dimensional mixed-type data are prevalent in a large cohort study. Our study is motivated by the Canadian Healthy Infant Longitudinal Development (CHILD) cohort study. The classic method to discover lifestyle patterns is to use principal component analysis (PCA) to reduce dimensionality and create a small subset of new variables. However, since our mixed-type data consist of both continuous and categorical variables, PCA is not applicable, and a novel method is needed to address this problem. Objectives: This study aims to define lifestyle patterns using multi-dimensional data from parent-reported questionnaires and to examine whether the lifestyle patterns are consistent as children grow.
Methods: In an attempt to describe lifestyle patterns using multi-dimensional mixed-type data, factor analysis of mixed data (FAMD) was used to reduce the data dimensionality and group continuous and categorical variables together to form lifestyle patterns. We also use correlation plots to observe the relationship between variables.
Results: For year 3 data, five dimensions are selected, and they account for 59% of the variation in the original data. For year 5 data, six dimensions are selected, and they account for 69% of the variation in the original data. Lifestyle Patterns are identified in each dimension.
Conclusion: Several different lifestyle patterns were found for children aged 3 and 5. One of them is consistent over time. We can associate these lifestyle patterns with common diseases in pre-school-aged children, such as obesity and asthma. However, due to the limitation of FAMD, further analyzes can be conducted with missing values imputation, and more lifestyle variables added.

Machine learning approaches for classifying credit card clients based on default risk

Mei Han

Introduction & objective: This project intends to compare several machine learning methods for the classification of clients regarding their risk of default based on information of default payments, demographic factors, credit data, history of payment, and bill statements. The database been used contains information about credit card clients in Taiwan from April 2005 to September 2005.
Methods: I mainly applied (1) logistic regression, (2) tree-based classification method and (3) SVM for this classification. Since the data set is unbalanced, I applied some methods for improving the classification accuracy, including (1) under-sampling and (2) ensemble techniques. And I use AUC for comparing the accuracy of classification.
Results: (1) Techniques like under sampling and ensemble tree-based model can improve the accuracy. (2) These models demonstrate similar AUC, and in all these approaches the information about the activity one month before the default plays an important role in predicting.
Conclusion: For the practice in the risk management of banks, we would like to check clients' payment activities frequently and set proper warning lines to take actions in advance, regarding different risk tolerance of different institutions.

Relationships between Major Health Behaviors and Sleep Problems: Results from 2017-2018 Canadian Community Health Survey

Mo Zhou, Rosane Nisenbaum

Introduction & Objectives: This cross-sectional study describes sleep duration and quality by gender in a population aged from 18 to 64 and investigates how health behaviors, including smoking, alcohol consumption, fruits and vegetable (FV) consumption and physical activity, are associated with short/long sleep duration and sleep quality problems (difficulty initiating/maintaining sleep [DIMS], finding sleep refreshing and daytime sleepiness).
Methods: Using Canadian Community Health Survey data from cycle 2017-2018, multinomial and binary logistic regression models were computed.
Results: Of the 36,431 respondents included, only 51.9% of respondents met the recommended sleep duration. 57.1% of females and 43.9% of males reported DIMS. Any form of binge drinking was associated with increased DIMS, with the highest odds being among with males reporting weekly binge drinking (odds ratio (OR) 1.74 [1.31,2.32]). Binge drinking was also associated with increased odds of short sleep among females only (OR 1.40 [1.10,1.76] in occasional binge drinking). Daily smokers had higher odds of short sleep (OR 1.35 [1.15,1.57], among females; similar OR among males) and lower odds of finding sleep refreshing (OR 0.74 [0.63,0.86] and OR 0.86 [0.73,1.00]; females and males, respectively). Similarly, former smokers had higher odds of daytime sleepiness (OR 1.53 [1.07,2.20]) and short sleep (OR 1.13 [1.00,1.28]) among females only. Increased FV consumption was associated with lower odds of DIMS (OR 0.92 [0.86,0.99]) for females.
Conclusion: There is a high prevalence of sleep problems among Canadians, which tend to be associated with the unhealthy behaviors. Also, gender differences in the relationships between health behaviors and sleep problems emerged.

Sex-stratified vs sex-combined analysis in the presence of genetic effect heterogeneity

Boxi Lin, Lei Sun

Introduction & Objective: The effect of a genetic variant on a complex trait may differ between male and female, e.g. genetic effects may be sex-specific for testosterone levels. In the presence of genetic effect heterogeneity between female and male, sex-stratified analysis is often used, which provides easy-to-interpret sex-specific effect size estimates. However, from power of association testing perspective, sex-stratified analysis may not be the best approach. As sex-specific genetic effect implies SNP-sex interaction effect, jointly testing SNP main and SNP-sex interaction effects may be more powerful than sex-stratified analysis or the standard main-effect testing approach.
Objectives: Our primary objective is to compare the power of sex-stratified and sex-combined testing approach in the presence of genetic effect heterogeneity. Our secondary objective is to study if the interaction analysis can be derived from sex-stratified summary statistics when individual data are not available.
Methods: We considered several different sex-combined methods and evaluated them through extensive simulation studies. And we provide additional supporting evidence by utilizing the publicly available sex-stratified GWAS summary statistics of testosterone levels of the UK Biobank data.
Results: We observed that a) the joint SNP main and SNP-sex interaction analysis is most robust to a wide range of genetic models, and b) this joint interaction testing result can be obtained by quadratically combining sex-stratified summary statistics (i.e. squared sex-stratified summary statistics).
Conclusion: We conclude that there is no uniformly most powerful test in various scenarios of genetic effect heterogeneity. However, genetic heterogeneity can be leveraged through gene-sex interaction analysis which gives a robust and powerful association test. In addition, joint test of genetic main and gene-sex interaction effect is empirically the same with meta-analysis that quadratically combines sex-stratified summary statistics.

Dealing with Partially Observed Confounders and Feature Selections in Propensity Score Analysis: A Comparison of Different Approaches with Applications in Non-alcoholic Fatty Liver Disease

Yun Zhu, Sareh Keshavarzi, Rahi Jain, Mamatha Bhat

Introduction & objective: Propensity score (PS) analysis is a popular approach that can reduce confounder effects on the treatments. The bias of the estimated exposure effect depends strongly on which confounders are included in the model. Moreover, the analysis will be complicated when the covariates used to estimate the PS are only partially observed. This study aimed to explore the impact of using different combination methods for feature selection (FS), multiple imputations (MI), and PS techniques on the treatment/exposure effect estimates and provide practical recommendations for future PS analysis.
Methods: A sub-cohort of clinical data of Non-alcoholic fatty Liver disease (NAFLD) information from UK Biobank data was used as our motivating example. Five different combination types of LASSO were used as our FS methods. Two MI approaches include combining the treatment effects after MI (MIte) and combining the PS after MI (MIps). Complete case analysis (CCA) was used to deal with the partially observed confounders. Matching, stratification, and inverse probability of treatment weighting (IPTW) were used for the PS approaches.
Results: Our results showed that the IPTW is the preferred approach for each FS method. In most cases, the MIps approach had better performance for dealing with missingness. For CCA, full matching outperforms other PS approaches, while for two MI methods, IPTW is better than others.
Conclusion: Different combination approaches may give different results depending on the data characteristics. In the future, we need to construct an automatic method that can find an appropriate pipeline from different combinations of these methods.

Childhood-onset Systemic Lupus Erythematosus: Defining Long-term Outcomes in Ontario

Ha-Seul Jeoung, Kuan Liu, Deborah Levy

Introduction & objective: Systemic lupus erythematosus (SLE) is an autoimmune disease that can cause damage to any organ in the body. It is characterized by chronic health conditions such as kidney disease. SLE that occurs before the age of 18 is known as childhood-onset SLE (cSLE). cSLE patients transition into adult care as they become older, which makes it challenging to conduct long-term follow-up studies. As a result, long-term effects of cSLE still remain unclear. The main goal of this project is to address this knowledge gap by examining long-term consequences of cSLE. Specific objectives include determining all-cause and cause-specific mortality rates and identifying baseline demographic and disease characteristics that are associated with higher risks of mortality and morbidities.
Methods: Approximately 600 cSLE patients were recruited from Ontario pediatric centers. Health administrative data containing long-term health outcome measures were examined through survival analysis methods. Kaplan-Meier estimates were used to calculate survival probabilities and multi-state survival model was fit to obtain hazard ratios.
Results: All-cause mortality rate was 3.33 per 1000 person-years. Baseline demographic characteristics including gender were not significantly associated with higher risks of mortality. The presence of psychosis was associated with lower survival probability. Increased number of SLE criteria was associated with higher rates of renal events. Other transitions did not have variables that were significantly associated with mortality or renal events.
Conclusion: We investigated the long-term effects of cSLE by examining mortality rates and hazard ratios. Such information will be beneficial for physicians when providing guidance to patients.

Feature selection with ElasticNet to assess relationship between sleep and depression treatment

Michelle Wu, Wendy Lou, Venkat Bhat, Sidney H Kennedy, Raymond Lam

Introduction & Objectives: Major Depressive Disorder is the leading cause of disability in the world, with the World Health Organization estimating that 5% of the total adult population is affected by it. There exist quantitative ways to measure depression via questionnaires and assessments. Seeing how the process for treatment can be difficult at times though, it is important to think if our assessment methods are capturing the most that they can. A key symptom of depression is sleep-related disturbances which may even act as an early predictive symptom. It is of interest to see if there are certain sleep-related disturbances that may better predict an individual's treatment response and if there are differences in sleep patterns between those who are responding to treatment and those fully in remission.
Methods: Data from the Canadian Biomarker Integration Network (CAN-BIND) will be utilized for feature selection. Using ElasticNet, a regularized regression method, we aim to detect specific sleep-related questions from the Pittsburg Sleep Quality Index (PSQI) that may be stronger predictors of treatment response. A patient's treatment response will be based on their questionnaire scores from the Montgomery–Asberg Depression Rating Scale (MADRS).
Results & Conclusion: Preliminary results indicate that a decrease in daytime dysfunction and improved sleep quality are predictors for both response and remission groups. Further investigation will need to be conducted to understand the results in a clinical context and to see if a consolidated group of sleep symptoms can be determined from the PSQI for prediction purposes.

Nursing Involvement in AI Research

Mark Germano, Alistair Johnson

Introdroduction & Objective: In critical care, nursing staff document approximately 800 items per hour. Yet, research in artificial intelligence rarely includes domain expertise from nursing staff. We take a deeper investigation into Nursing involvement in AI research by analyzing the authors of research papers.
Methods: We reviewed 258 papers that covered machine learning within Intensive Care Units (ICU). A topic model was created on each abstract to categorize papers. This model allows us to understand the nature of our dataset as well as areas in which certain author degrees may be more common. Authors from the 258 papers were then extracted along with their education titles, affiliations, and current country location. We used Python to scrape online databases (Scopus and ORCID) with documented author information. Any missing information was then manually input to complete the dataset. With the author information, we plan to perform a network analysis that will help to summarize the papers based on education titles of authors. Additionally, meaningful subgroups will be generated based on other author information that was documented (this may include things like locations or institution).
Results: It was found that main categories of papers were Severity of Illness/Acute Physiology Score, Machine Learning/Predictive Modeling, False Alarms, Neonatal ICU, Sepsis, and Electroencephalography. Based on preliminary analysis, it appears that Nursing involvement is minimal as initially suspected.
Conclusion: AI research in health care is a growing field that makes use of data. Nurses are prevalent in Hospitals while also documenting very valuable data, however appear to have minimal involvement in AI research.

Multiple Imputation Methods for Multilevel Ordinal Outcomes

Mei Dong, Aya Mitani

Introduction and objectives: Multiple imputation (MI) is an established technique to handle missing data in observational studies. Joint modeling (JM) and fully conditional specification (FCS) are commonly used methods for imputing multilevel clustered data. However, MI methods for ordinal clustered outcome variables have not been well studied, especially when there is informative cluster size (ICS). The purpose of this study is to describe different imputation strategies for the multilevel ordinal outcome when ICS exists.
Methods: We conducted comprehensive simulation studies to compare five different methods: complete case analysis (CCA), FCS, FCS+N (include cluster size when performing the imputation), JM, and JM+N under different scenarios. Cluster weighted GEE, a marginal analysis model to deal with ICS, was implemented to fit the imputed dataset. We further applied these methods to a real dental study.
Results: The mean relative biases for intercept eta1 are 11.53% from FCS+N, 19.78% from FCS, 19.62% from JOMO+N, and 26.71% from JOMO, followed by CCA with 76.03% when both ICS and interclass correlation coefficient are large and data is missing at random (MAR). When missing not at random (MNAR), all methods are highly biased. However, FCS still has better performance than JM and CCA. The real data analysis shows that patients with Metabolic syndrome have lower odds of having healthier CAL scores (OR=0.81 based on FCS+N).
Conclusion: Simulation results show that including cluster size in the imputation can significantly improve imputation accuracy when ICS exists. FCS provides a more accurate and robust estimation than JM, followed by CCA.

Abstracts Session 2

Thursday, May 12th - 14:00 - 16:00 EDT, Chair: Tony Panzarella

Click here for Zoom link

Identifying characteristics and strategies for long-term success in the management of pediatric systemic lupus erythematosus

Fatema Tuj Johara, Kuan Liu, Linda Hiraki

Introduction & objectives: Systemic lupus erythematosus (SLE) is a rare, chronic, and autoimmune disease. It has been shown that patients with childhood-onset SLE have more aggressive disease course and organ damage than patients with adult-onset SLE. In this project, we were interested to examine disease activity and damage of the patients with childhood-onset SLE. Particularly, we aimed to identify factors and the patient group that is associated with the risk of damage (1) at graduation (at 18 years) and (2) for the first time of life.
Methods: A total of 488 patients who were newly diagnosed with lupus at SickKids Lupus Clinic between the years 1985 and 2016 were included in this prospective cohort study. The patient's baseline and clinical data were collected at subsequent follow-up visits. Univariate and multivariable logistic models were applied to identify factors associated with damage at graduation. Univariate Cox models were considered for examining risk factors at first damage.
Results: Of 488 patients, 82% were females, the median age of diagnosis was 14.5 years, and the majority of self-reported ethnicity group was European (32%). Statistical analysis showed that patients who have been diagnosed with lupus in earlier years were more at risk of damage. South Asian ethnicity group was at more risk compared to the Europeans. People who took a higher dose of steroids as well as had the presence of some SLE features were at greater risk of organ damage.
Conclusion: Recognizing patients' groups and risk factors of organ damage might help design a targeted treatment policy.

Integrating Genetic & Clinical Risk Factors on UKBiobank data in predicting NAFLD Risk using Machine Learning Algorithm

Jiayin Chen, Sareh Keshavarzi, Divya Sharma, Mamatha Bhat

Introduction: Nonalcoholic fatty liver disease (NAFLD) is one of the most prevalent liver diseases in Canada. The aim of our study was to analyze the impact of individual genetic variants on NAFLD and develop a machine learning (ML) algorithm integrating clinical, demographic, and genetic risk factors to predict NAFLD risk.
Methods: We selected a cohort of patients from the UK Biobank white British ancestry. NAFLD status is diagnosed according to ICD10. A total of 2088 patients with NAFLD constituted cases and 428,064 subjects without NAFLD constituted controls after quality control.
Backward selection method is applied in Logistic regression to analyze the impact of individual variants on NAFLD. Furthermore, 6 different ML approaches (Logistic regression, Ridge, Lasso, Gradient Boosting, RandomForest, CatBoost) integrating clinical, demographic, and genetic variables are evaluated for predicting NAFLD. Cross validation and receiver operating curve are used for validation and comparison.
Results: In genetic impact, single-nucleotide polymorphisms (SNPs) in HSD17B13 and PNPLA3 gene were most significant, which encodes an enzyme in hepatocytes, increases the whole spectrum of liver damage related to NAFLD respectively. The ensemble-based ML model achieved an area under the curve of 0.831 using the Gradient Boosting model. The most significant clinical variables observed by the best predictive modeling were alanine transaminase (ALT), body mass index and Glutamyl Transferase (GLT).
Conclusion: Our study suggests that though each SNP makes little contribution to the prediction, integrating genotypes train new model can slightly increase AUC in each model (2%). Lifestyle habits have a more important influence than genetic effects. Patients with higher ALT, GLT and BMI should be flagged as NAFLD cases for screening.

A Comparative Study of Robust Estimation Methods for Cox Regression Model

Shiyao Ying, Xuan Li, Amy Liu

Introduction & Objective: Cox proportional hazard model is a semi-parametric model that is widely used for estimating hazard ratio in survival analysis. Previous studies have shown that mild departures from the model may lead to inaccurate and even incorrect inferences. This study compares several robust estimation methods that address the issue of outliers in cox models.
Methods: Methods including robust estimator by including weights on observations, using concordance C-index to do the stepwise deletion and bootstrapping, and excluding outliers by deviance residuals were compared with simulations. Two types of outliers, geometric and probabilistic outliers were assumed in simulation scenarios. Different censoring types, censoring percentages, outlier percentages, and sample sizes were also considered to vary in the study. Real data analyses were conducted on a breast cancer dataset and a myeloma dataset to check the coefficient change and common outliers selected by the robust estimation methods.
Results: For geometric outliers, which lay far away from the center of the data cloud, concordance C-index-based methods including stepwise deletion and bootstrapping hypothesis test perform better, especially for larger sample sizes (n=100). For probabilistic outliers, which are from different distributions, robust estimation methods by adding weights performed relatively better than concordance-based methods. The real data analysis suggested that common outliers selected by deviance residuals and concordance C-index-based methods have relatively longer or shorter time to event than most subjects.
Conclusion: In general, all these methods trying to address the outlier problem did a decent job and we can select them based on our scenarios.

Identifying informative subsample selection strategies for shotgun metagenomics analysis in microbial studies

Daniel Felipe Segura Hincapie, Divya Sharma, Osvaldo Espin-Garcia

Introduction & Objective: The microbiome is increasingly regarded as a key component of human health, and analysis of microbiome data can aid in the development of precision medicine. Due to the high cost of shotgun metagenomic sequencing (SM-seq), microbiome analyses can be done cost-effectively in two phases: Phase 1-sequencing of 16S ribosomal RNA, and Phase 2-SM-seq the microbiome of a subsample. Existing research suggests strategies to select the subsample based on biological diversity and dissimilarity metrics calculated using operational taxonomic units (OTUs). However, the microbiome field has progressed towards amplicon sequencing variants (ASVs), as they provide more precise microbe identification and sample diversity information. The project objectives were to compare the subsampling strategies for two-stage metagenomic studies when using ASVs instead of OTUs, and to propose data driven strategies for subsample selection through dimension reduction techniques.
Methods: We used 199 samples of infant-gut microbiome data to generate ASVs and OTUs, then generated subsamples based on existing subsampling methods. Additionally, we utilized principal component analysis of the ASVs and OTUs for dimension reduction and subsample selection. Linear discriminant analysis Effect Size (LEfSe) was used to assess differential representation of taxa between the subsamples and total sample for each method.
Results: Among the pre-established subsampling methods used, there was 40-91% agreement in the subsample selection between OTUs and ASVs, and the bacterial representation was similar across all methods.
Conclusion: Each subsampling methodology yields similar results for ASVs and OTUs. We are also assessing the effect of each subsampling method on downstream SM-seq.

Investigating the association between military length of service and rate of emergency department visits following release

James Saunders, Rinku Sutradhar, Aitken Alice, Cramm Heidi, Mahar Alyson

Introduction & Objectives: There is limited research available on Canadian Armed Forces (CAF) Veterans use of emergency department (ED) services, relative to non-Veterans. Understanding ED usage can provide insight into primary care access, physician attachment and other needs of military Veterans. Our aim was to compare rates of ED visits between Veterans and non-Veterans, and examine the effects of sex and length of service on these rates.
Methods: This is a retrospective, matched cohort study of Ontario CAF and RCMP Veterans and non-Veterans, developed from administrative databases housed at the Institute for Clinical Evaluative Sciences. Veterans residing in Ontario who were released between January 1, 1990 and March 31, 2013 were eligible for study inclusion, and were matched to four civilian comparators.
Results: The crude ED visit rates of Veterans and matched civilian comparators were 3.20 (3.18-3.23) and 3.15 (3.13-3.16) per 10 person-years, respectively, with a crude relative rate of 1.02 (0.90-1.16). The adjusted relative rate was 0.96 (0.93-0.98). An interaction between sex and Veteran status was significant. Length of service was inversely associated with ED visitation rate, with relative adjusted ED visit rates ranging from 1.17 (1.09-1.26) to 0.78 (0.75-0.82) for Veterans with <5 and $geq30$ years of service, respectively.
Conclusions: Veterans had a lower rate of ED utilization compared to the general population. The association between ED visit rates for Veterans and non-Veterans varied by sex and length of service. This could signal differences in underlying acute health needs or access to healthcare following release.

Workflow for Identifying Protein of Interest

Jiayue (Irene) Feng, Lisa Avery

Introduction & objective: Mass Spectrometry can be used to measure the intensity of proteins in tissue from the central nervous system. Of interest is whether differences in protein levels can differentiate people with MS for non-healthy group from healthy controls. Thousands of proteins are often identified for each participant, so the number of participants is small relative to the number of potential proteins of interest. The target of this project is a reproducible workflow for evaluating the protein intensity measures and identifying the individual proteins that collectively identify MS. We also want to see if predictor age and sex influence our findings.
Methods: For analyzing the above question, I apply t-tests for screening the reliable proteins and fit them into logistic regression model for further protein screening.
Results: Our results show that there are 129 proteins that are reliable detected after screening. The factors age and sex have no effect between the two group when we screen the proteins.
Conclusion: Among those proteins, most of them are highly discriminated between groups. Overall, age and sex do not have effect on our result by looking at their p-values. The log-transformation is necessary since it has lower p-value compared to the original data when I was performing t-test.

A Gaussian mixture model-based variational graph autoencoder algorithm for clustering single-cell RNA-seq data

Eric Lin, Pingzhao Hu, Leann Lac, Boyuan Liu, Daryl L.X Fung

Background: Cell type identification from single-cell RNA sequencing (scRNA-seq) is crucial to understand disease mechanisms for disease diagnosis and drug discovery, which involves classifying the data into clusters of single cells.
Objectives: scRNA-seq data is high dimensional with many analysis challenges. In this study, we propose to integrate advanced statistical modeling and deep learning and develop a Gaussian mixture model-based variational graph autoencoder (GMM-VGAE) to address this important issue.
Methods: GMM-VGAE is an unsupervised clustering algorithm that takes inputs of cell-cell graph and gene feature matrix to feed through the graph encoder and decoder to reconstruct the cell-cell graph. We applied the method and other three baseline models to cluster the single cells from three publicly available datasets.
Results: Using adjusted Rand index (ARI) as model performance measure, we showed the GMM-VGAE algorithm has better performance than other baseline models in all three datasets. With ARI of 0.940 for the Baron3 dataset, 0.948 for Baron4 dataset, and 0.936 for Darmanis dataset.
Conclusion: Combination of statistical modeling and unsupervised deep learning algorithm can be successfully applied to analyze scRNA-seq data. By incorporating GMMVGAE algorithm in scRNA sequencing, the performance of future analysis and application can be enhanced.

Assessing longitudinal trends in online brain health assessment scores for older adults living in community settings

Luis Ledesma, Sandra Gardner, Malcolm Binns, Larissa McKetton

Introduction & Objectives: There is an increasing need for reliable Internet-based screening tools for cognitive assessment in middle-aged and older adults. The Cogniciti Brain Health Assessment (BHA) tool consists of four online tasks that are sensitive to early signs of cognitive decline; and a questionnaire that records participants' information. The test-takers were recruited through word-of mouth, media outlets, and advertisements. In our analysis, we consider a sample of 1980 adults from the longitudinal dataset, where an individual has 2 to 4 assessments available, with different follow-up times between assessments. In particular, we are interested in determining whether the test results show evidence of cognitive decline over time, as previous studies have mainly focused on the use of cross-sectional samples.
Methods: As standard longitudinal methods may not be entirely applicable due to the presence of irregular assessments, visit processes, and their assumptions are introduced. Assuming that performance at previous assessment may influence whether or not an individual takes future tests, the outcome and assessment intensity are modelled through inverse-intensity weighted generalized estimating equations (IIW-GEE).
Results: Compared to linear mixed models which gave singular boundary fits, the IIW-GEE method was able to obtain marginal model estimates for all four outcome tasks. Using the IIW-GEE framework, our findings are that subjects with memory concerns and a higher age at baseline had worse predicted baseline performance in the shape match task. In addition, those who reported health concerns and a high school (or lower) level of education ended up having a lower assessment intensity.
Conclusion: IIW-GEE methods allow us to model the assessment intensity as a function of participants' performance on the previous assessment and baseline covariates, where we are able to account for the fact that test-takers see their percentile score after taking the assessment.

Evaluate the effect of data imputation on clustering analysis of single-cell RNA sequencing data

Boyuan Liu, Eric Lin, Pingzhao Hu

Background & Objectives: The key challenge in analyzing scRNA-seq data is the large proportion of missing values (zeros) due to the fraction of non-sequenced transcripts. It creates a need for imputing the missing values. However, previous studies showed that most imputation methods could not distinguish the biological zeros and technical zeros and disrupted the intrinsic data structure during imputation, which disimproved the performance of traditional downstream clustering analysis. In recent years, the novel deep learning-based clustering method GMM-VGAE has been developed to be insensitive to the missing values in scRNA-seq data and outperforms the traditional methods. This study aims to evaluate whether data imputation affects the novel downstream scRNA-seq clustering method; If it does, what data imputation methods should be applied.
Methods: Two state-of-the-art imputation methods (SAVER and scImpute) will be applied to impute the missing values in three scRNA-seq datasets with cell labels (ground truths). Both methods are based on probabilistic models and able to impute the technical zeros while preserving the biological zeros. Four unsupervised clustering methods (GMM-VGAE, SAM, scVI, and CellVGAE) will be applied to cluster the cells. The clustering results will be evaluated by the ARI metric and compared to the baseline clustering without imputation. The analyses will also be performed using the simulated data from the scRNA-seq simulator SPARSim.
Expected results: Related scRNA-seq datasets have been collected and preprocessed. Imputation methods, clustering methods, and performance metrics have been selected, and related Python libraries and R packages have been identified. The data analyses are ongoing.

Using classification and regression trees to model missingness in youth BMI, height, and body mass

Amanda Doggett, Ashok Chaurasia, Jean-Philippe Chaput, Scott T. Leatherdale

Introduction & Objectives: Research suggests that there is often a high degree of missingness in self-reported body mass index (BMI) data among youth, which may bias research findings. Although public health relies on these data for population surveillance of body adiposity as well as conducting research surrounding overweight and obesity, very few studies have focused on examining missingness and related bias in this domain.
Methods: This study used classification and regression trees (CART) to examine missingness in youth height, body mass, and BMI among 74 501 youth who participated in the COMPASS study in 2018/19. COMPASS is a longitudinal cohort study of youth in Canada which examines a variety of diet, movement, academic, mental health, and substance use variables; several variables in these domains were included in the CART models to examine their associations with missing BMI, height, and body mass.
Results: Findings suggest that social desirability played a large role in nonreporting among both males and females in this sample, and that those who perceived themselves as overweight were more likely to be missing BMI. This mechanism of missingness may have important implications for research which uses similar youth self-report height and weight data.
Conclusion: This study adds to the limited existing research on missing height, body mass, and BMI data among youth, and identifies potential bias from nonreporting. This study also demonstrates the utility of CART models for examining missingness, highlighting how they may be used as an initial step to the appropriate handling missing data.

Identifying Genetic Determinants of Absence Seizure Incidence in Juvenile Myoclonic Epilepsy

Eric Sanders, Lisa Strug, Naim Panjwani

Introduction & Objective: Component traits of Juvenile Myoclonic Epilepsy can vary, and recent research has highlighted an association between absence seizure incidence and anti-seizure drug resistance. This study aims to identify genetic determinants of absence seizure incidence, which would improve understanding of disease etiology and potentially lead to the development of novel treatments with improved outcomes.
Methods: A genome-wide association study was completed, testing 9.3 million SNPs across 23 chromosomes for association with absence seizure incidence, using a logistic regression framework controlling for sex, population strata, and genotyping cohort, and allowing for genotype-sex interaction. For each SNP, the test of interest was a score test with two degrees of freedom, testing for genotype effect in at least one sex.
Results: Preliminary results suggest that a genetic region on chromosome 17q11.2 has a potential association with absence seizures that differs by sex. The observed lead SNP in this region was rs75671533 (C/T) with T allele frequency 16.6%. We observe a protective effect in females (OR=0.31, 99% CI=0.273,0.875) and opposite or neutral effect in males (OR=1.85, 99% CI=0.924,3.704) (with a significant two degrees of freedom score test p-value of 2.48 X 10-8). The implicated genetic region is intergenic.
Conclusion: A genetic region has been identified as being associated with absence seizure incidence, with differing effect in males and females. There is a need for further study to identify potential mechanisms of action that would explain the association with absence seizure incidence. Future work will include colocalization analysis to identify any influence the SNP may have on gene expression.

Assessing the Impacts of Serial Correlation on the Performance of Structural Change Detections Using Simulation Method

Zixin Zhou, Rahim Moineddin

Introduction & Objectives: Structural change detection is crucial in evaluating the longitudinal effects of educational, administrative or policy interventions. Methods are mainly based on the F test and generalized fluctuation test framework. In time series, autocorrelation is very common, but how it affects the results of structural change detections is under discussion. Thus the objective is to explore the performance of these three tests evaluated by type I error in the presence of MA(1) autocorrelated regression error.
Method: The methods for structural change detection applied in this project are the Chow test, segmented regression analysis and Cumulative sums of recursive residuals. The data are simulated from a simple linear regression model where error terms associated with each observation followed a MA(1) autocorrelation process. Fitted models for the Chow test and Cumulative sums of recursive residuals analysis are the same and these two methods assume that the error terms are uncorrelated while Segmented Regression rests on the assumption of independence that it allows autocorrelation in regression errors and adjusts parameter estimations through PROC AUTOREG in SAS.
Results: All three methods show that when regression errors are positively autocorrelated, the type 1 error increases as autocorrelation gets stronger. When regression errors are negatively autocorrelated, the type 1 error decreases as autocorrelation gets stronger.
Conclusion: Results of these three methods are not reliable in the presence of autocorrelation, even though the Segmented Regression incorporated autocorrelation. There is high inflation/deflation in type 1 error. So structural changes tend to be missed when negative autocorrelation exists and researchers are biased toward detecting structural change when it actually does not exist in the presence of high positive autocorrelation. In this case, it is highly possible for interventions to be overestimated.

Are Vision and Gait Associated with Fear of Falling Among Patients with Neurodegenerative Disease?

Lee Radigan, Malcolm Binns

Introduction/Objectives: Falling is an issue for the elderly that can lead to poor health outcomes. Factors associated with falling include vision and gait impairments. Fear of falling (FoF) exists even for individuals who have not yet experienced falls. The purpose of this study was to investigate whether both vision and gait are associated with FoF among patients with neurodegenerative disease.
Methods: Clinical, gait and vision data collected by the Ontario Neurodegenerative Disease Research Initiative, with a focus on 55 participants from the cohort of patients diagnosed with Alzheimer's disease (AD) or mild cognitive impairments (MCI) were used. The outcome of interest is fear of falling (FoF; 0-10 Likert). Due to a high number of correlated variables, the Elastic Net penalized regression technique was applied. An optimal model as well as competing models were compared. Additionally, "gait-only" and a "vision-only" models were examined.
Results: The model with the minimum root mean squared error (RMSE) was selected. Competing models with similar RMSEs were also observed. The number of non-zero coefficients remaining in the selected model and the two competing models ranged from 15 to 17. When removing gait measures from the analysis, the variables selected for our model remained relatively unchanged. Conversely, when removing vision variables, new non-zero coefficients appeared that were not present in the selected model.
Conclusion: It is important to look at a range of models before interpreting model coefficients from elastic net. The observed range uncover differences in the models which can lead to challenging interpretation. Working closely with clinicians is necessary in order to properly scrutinize the model coefficients.

Bayesian Cox Proportional Regression with Partial Likelihood

Ziang Zhang, Alex Stringer, Patrick Brown, James Stafford

Introduction & Objective: For problems involving time-to-event data, the combination of Cox proportional hazard (Cox PH) models and inference via partial likelihood has been the dominant methodology following its development by Cox. The inference that is conducted via partial likelihood does not require assumptions to be made about the form of the baseline hazard. Further, the use of Bayesian inference with the Cox PH model is desirable as this yields model-based estimation and uncertainty quantification for all parameters of interest in the presence of complex models for the hazard, which would be difficult to achieve otherwise.
Method: We propose a flexible and scalable approximate Bayesian inference methodology for the Cox Proportional Hazards model with partial likelihood. The model we consider includes semi-parametric covariate effects and correlated survival times. The proposed method is based on nested approximations and adaptive quadrature, and the computational burden of working with the log-partial likelihood is mitigated through automatic differentiation and Laplace approximation.
Result: We demonstrate the practical utility of our method and its computational advantages over Markov Chain Monte Carlo (MCMC) methods through the analysis of kidney infection times, which are paired, and the analysis of Leukemia survival times with a semi-parametric covariate effect and spatial variation.
Conclusion: The methodology we proposed provides a flexible way to carry out Bayesian inference for Cox proportional hazard models with partial likelihood, that accommodates the inference for semi-parametric covariate effects, spatial variations, and correlated survival times.

COVID-19 public health measure adherence amongst parents and children: investigating the importance of sociodemographic factors during lockdowns and reopenings

Kevin Dang, Charles D.G. Keown-Stoneman

Introduction & Objectives: This longitudinal cohort study of parents and children was conducted in the Greater Toronto Area from April 2020 to present day. The primary outcome of this study was total adherence to COVID 19 public health measures in the past seven days, and the primary exposure was calendar date (and lockdowns/school closures) when the parents completed the weekly data collection form. To investigate how the adherence to COVID 19 public health measures amongst parents and children in Ontario changes over time, whether provincial lockdowns were associated with higher adherence to public health measures amongst parents, whether school closures were associated with higher adherence to public health measures amongst children, and which sociodemographic factors modified these associations.
Methods: Piecewise linear mixed effects regression to test for changes in adherence during lockdowns, school closures and reopenings. 10-fold cross-validation was used to select from the 7 sets of interaction terms in the models for parents and children.
Results: For children, all 7 sets of interaction terms were selected, compared to 4 sets for parents. On average, total adherence to the five public health measures decreased over time for both parents and children. Additionally, provincial lockdowns were associated with higher adherence to public health measures amongst parents and school closures were associated with higher adherence to public health measures amongst children. For sociodemographic factors, the result varied depending on the closure or open period.
Conclusion: Overall, total adherence decreased over time with exceptions of temporary jumps during lockdowns and school closures. There were no significant differences within many sociodemographic factors at most time periods. For future analyses, other factors should be investigated.

A machine learning approach to differentiate between COVID-19 and influenza infection using synthetic infection and immune response data

Suzan Farhang-Sardroodi, Jane M Heffernan, Morgan Craig

Introduction: Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and influenza viruses cause COVID-19 and influenza diseases, respectively, and mainly infect the upper and lower respiratory tract. Both infections present similar primary symptoms such as cough, fever, sore throat, runny or stuffy nose, tiredness, and body aches. Early on in infection this can lead to a clinical dilemma in diagnosis. While influenza and COVID-19 have the same primary signs, COVID-19 can produce more severe symptoms, illnesses, and higher mortality. Recently, COVID-19 has, through its worse overall decompensation due to its intensive transmission and vascular effects, caused an unrivaled global crisis. As the globe moves to endemicity, as the striking COVID-19 outbreak continues, the concurrence of COVID-19 and influenza epidemics is impending.
Objective: We aimed to design a data analysis tool that not only can accurately differentiate between various infections but also could be applied as a low-cost classification system that would not require expensive virus typing procedures and could rely solely on viral load and interferon measurements.
Method: Herein, we used a simple ML-based classification to rapidly identify patients with influenza or SARS-CoV-2 based on the main features of the within-host viral dynamics and the immune response. Our model employed a Lasso regression classifier trained to externally identify between at least two hundred virtual patients.
Results: To identify between COVID-19 and influenza patients, our model ascertained a good performance when the data distributed during the main infection period (ROC AUC = 95%) or even for the early days of infection after the incubation period (ROC AUC = 91%). Analyzing the feature importance revealed that the viral load and the productively infected cells are the most important components to determine if a patient is infected by influenza or SARS-CoV-2.
Conclusion: The current investigation highlighted the ability of machine learning models to accurately identify two different diseases based on major components of viral infection and immune response. Our model was trained and successfully evaluated on synthetic data. The model, however, could be applied to animal or human clinical data. This could be useful, for example, if a clinical trial is complicated by the existence of an infectious disease with similar infection characteristics.

Machine Learning for Radiologic Series Categorizations of CT Scan

Suvd Zulbayar, Ervin Sejdic, Errol Colak

Introduction: CT scans are composed of a series of images that are related by body part, kernel, axis, phase type, and IV/oral/rectal contrast presence. Our goal is to develop machine learning models that classify each of these multi-class/binary parameters from CT scan images.
Method: We have 2622 labeled CT scans ranging from 13 to 1998 images in each series. A series was represented by slices of images to reduce computational complexity. The dataset is split as 80% for training, 10% for validation, and 10% for testing. Convolutional Neural Network (CNN) architecture called VGG16 was implemented for image classification. Stochastic gradient descent with momentum was used for optimization.
Results: For body part classification, 25th-50th-75th quartile slices were used. Class distribution is imbalanced, however, there is no highly dominant label, so CNN was able to learn. The best performing model was trained for 6 epochs with a learning rate of 0.001 and a batch size of 1. Test set accuracy was 84.48%. Kernel classification had improved performance on an artificially balanced dataset. The best model was trained for 11 epochs with a learning rate of 0.0001 and a batch size of 32. Test set accuracy was 83.91%. IV contrast classification had a better performance on an artificially balanced dataset. The best model had 94 epochs, a learning rate of 0.0000001 with momentum=0.99, and a batch size of 32. Test set AUC was 0.886.
Conclusion: Model performances show promising results especially considering the reduction of the dataset from 3D to 2D tensor, which could be further improved with more labeled data and a more balanced dataset.

Maybe we should stay home: Incorporating behavioural change into epidemic models

Madeline A. Ward, Lorna E. Deeth, Rob Deardon, Caitlin E. Ward

Introduction and Objective: As we have observed throughout the COVID-19 pandemic, behaviour often changes based on the current perceived risk of contracting the disease. In turn, this behaviour change can have a large impact on the transmission dynamics of the disease. While behavioural change has previously been incorporated into mathematical disease models, it has yet to be thoroughly explored in statistical models. The objective of this project is to explore a new class of statistical models that can incorporate the effect of behaviour change on infectious disease transmission.
Methods:Individual-level models can flexibly incorporate information on individual risk factors, including spatial location. This can account for the high degree of heterogeneity that is characteristic of population mixing, and, thus, infection transmission. However, these models have typically assumed stable population behaviour over time. This poster will present a new class of "behavioural change individual-level models" (BC-ILMs) where various functions of infection prevalence affect susceptibility or population mixing and illustrate their use through a simulation study.
Results: The results of the simulation study demonstrate that BC-ILMs can produce realistic patterns of disease transmission, including multiple epidemic peaks. They also show that models that do not include a behavioural change effect are not adequate in scenarios where behavioural change is present. However, BC-ILMs can produce accurate results even when there is no behavioural change in the data set.
Conclusion: Incorporating behavioural change into ILMs adds flexibility that may allow for more accurate epidemic response planning and forecasting.

A Comparison of Methods for Bayesian Inference in Clinical Trials

Ziming (Jocelyn) Chen, Anna Heath

Introduction & Objectives: Analysis using Bayesian methods updates inference as more data becomes available. Simulation-based approaches such as Markov Chain Monte Carlo(MCMC) are commonly used in Bayesian inference. However, there are barriers when applying Bayesian methods in practical settings because of the software complexity and high computational cost. A less computationally expensive approximation method, INLA which does not require simulations is available. The goal of the study is to compare INLA and two MCMC algorithms (in the software JAGS and STAN) in terms of method feasibility and estimation accuracy using clinical trial data.
Methods: The algorithms were compared using ATTACC and ACTIV-4a international Bayesian adaptive trial data that investigates the treatment effect of therapeutic anticoagulation with heparin in non-critically ill patients with covid 19 compared to usual care. By fitting Bayesian hierarchical generalized mixed models with categorical, binary, and time-to-event outcomes using JAGS, STAN and INLA, the posterior distributions of mainly the treatment effect were compared.
Results: INLA requires noticeably less computational time compared to STAN and JAGS (seconds compared to hours). All the 95% CIs for the treatment effect estimated using INLA overlapped with the simulation-based methods and the density curves for the posterior distribution of treatment effect almost overlapped for all three algorithms. One drawback of INLA is that it does not estimate the posteriors of the hyperparameters well.
Conclusion: INLA has higher feasibility of implementation with well established packages in statistical software like R and approximates the posterior densities of lower level of the hierarchy relatively well.

Big Data Clustering of SARS-CoV-2 Spike Glycoprotein Sequences

Vadim Tyuryaev, Jane Heffernan, Hanna Jankowski, Steven Wang, Derek Wilson

Introduction and objective: More than two years ago COVID-19 was declared as a global pandemic by the World Health Organization (WHO). Despite unprecedented public health measures and world-wide vaccination efforts, the pandemic is ongoing. Emergence of variants of concern such as Omicron further complicates the situation due to Omicron's increased transmissibility and immune escape capabilities. A plethora of studies attempted to uncover SARS-CoV-2 evolution from biological and medical perspectives, but only a few applied computational methods to study genetic similarities and differences among viral strains. We aspire to close the gap between biological and computational aspects by taking advantage of the abundance of high-quality SARS-CoV-2 genomic data through publicly available NCBI and GISAID databases.
Methods: We used a unique approach combining statistics and machine learning to meet the objectives. By utilizing such concepts as Entropy, Gaussian Mixture Models, Gower Distance, Partition Around Medoids we achieve a significant dimensionality reduction making our computations fast. We introduce the term "mutation-informed clustering" and show how such clustering allows for a noticeable separation among clusters of SARS-CoV-2 spike glycoprotein sequences. We also show how our approach can be used to predict new viral strains and to approximate the timing of their arrival.
Results: Our results point towards the existence of six global clusters of SARS-CoV-2 strains. By analyzing entropy changes, we are able to predict potential mutations 1-3 months in advance.
Conclusion: The COVID-19 pandemic created a unique situation in terms of sequencing data generated. We emphasize that such data should be fully utilized.

Deidentification of Free-text Clinical Notes

Peiqing Yu, Alistair Johnson

Introduction & Objectives: Deidentification refers to the removal of protected health information that can identify an individual. This task is challenging but crucial to protecting the patients' privacy during research while allowing a wider circulation of clinical data for the advancement of medical research. Our objective was to investigate the performance of a state-of-art method for deidentifying free-text clinical notes.
Methods: The transformers package developed by Hugging Face Co. was used in the study. The Bidirectional Encoder Representations model (BERT) architecture from the transformers package was adopted to do the named entity recognition task. The BERT model and its variations are already pre-trained and then we fine-tuned the model using our specific free-text clinical notes dataset. This paper fine-tuned the BERT and DistilBERT models first on i2b2 challenge datasets and then on the radiology reports clinical dataset. This study also assessed the performance of models on the training sets and datasets external to their training sets.
Results: The performance of the models was assessed by computing the precision, recall, and F1 measure. The F1 measures of the best performance model with fine-tuned hyperparameters can be greater than 95% and have a range of 78% to 95% for all protected health identifiers in the i2b2 2014 dataset. We also focused on the error analysis which is due to the massive amount of missing PHIs.
Conclusion: We can fine-tune the Bidirectional Encoder Representations model to achieve a state-of-art performance of deidentifying protected health identifiers on specific research datasets. This model architecture has the potential to be generalized to other corpora across the domains.

SSC Accreditation

Statistical Society of Canada (SSC) *offers two levels of accreditation:*

Professional Statistician (P.Stat.)
Associate Statistician (A.Stat.)

Why should I seek accreditation?

- *An ongoing professional development*
- *Membership in the national professional society*
- *Access to resources and advice from other statisticians for new statistical knowledge*
- *Mentorship program*

How to apply?

https://ssc.ca/en/accrediation

Société statistique du Canada (SSC) *offre deux niveaux d'accréditation :*

Statisticien professionnel (P.Stat.)
Statisticien associé (A.Stat.)

Pourquoi demander l'accréditation?

- *pour une exigence de perfectionnement professionnel*
- *pour appartenir à la société nationale professionnelle*
- *pour un accès à des ressources et des conseils d'autres statisticiens pour un développement des connaissances statistiques*
- *pour le programme de mentorat*

Comment s'appliquer pour l'accréditation?

https://ssc.ca/fr/accrediation

For practice in/ Pour pratiquer au Canada

Questions?

accreditation@ssc.ca

(613) 733-2662

Statistical Society of Canada
210-1725 St. Laurent Blvd.
Ottawa, ON K1G 3V4