SORA-TABA-DLSPH RESEARCH DAY

Virtual Poster Presentations

University of Toronto, Ontario
June 17th & 18th, 2020

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

SickKids Research Institute

McDougall Scientific — Insights you can trust

Contents

Welcome 2
Dedication 3
Organizing Committee 4
Professor Emeritus Paul Corey 5
Judges 6
Presentation Schedule Session 1 7
Presentation Schedule Session 2 8
Presentation Schedule Session 3 9
Abstracts Session 1 10
Abstracts Session 2 25
Abstracts Session 3 40

Welcome

On behalf of the Organizing Committee it is my distinct pleasure to welcome you to the Annual SORA-TABA-DLSPH 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, because of the COVID pandemic, both the workshop and the poster presentations will be presented virtually. The DLSPH Biostatistics Research Day Poster Presentations will take place over three separate sessions on June 17th and June 18th. This will be followed by the SORA-TABA Workshop on Statistical Machine Learning, featuring Dr. Noah Simon from the University of Washington, which is being planned for the fall.

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 cash 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 this year by the COVID pandemic. Keep well and stay safe!

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 in memory of Professor Peggy Ng

An accomplished statistician, Professor Ng first joined the department of Mathematics and Statistics at York University in 1996. She later joined the Atkinson Faculty of Liberal and Professional Studies, where she taught in the School of Administrative Studies, as professor of Applied Statistics and Decision Sciences. She served with dedication as director of the school from July 2011 until October 2015.

Professor Ng was devoted to increasing public awareness of the value of statistics and statistical thinking. She was past president of the Southern Regional Association of the prestigious Statistical Society of Canada, a forum for the exchange of ideas, methodologies, and practices between those in the field.

She received her PhD in Preventive Medicine and Biostatistics from University of Toronto in 1990. After joining York, she continued her research on biostatistics, experimental design, and psychometrics and their applications in health sciences.

Professor Ng leaves behind a rich legacy of research in the areas of patient safety and modelling of health statistics, in addition to countless papers on qualitative inquiries concerning environmental health, and quality of life in oncology.

Professor Ng died on Saturday, May 4, 2019 after a long illness. She was 64. She will be deeply missed by faculty, staff, and students.

Excerpt from https://laps.yorku.ca/2019/05/york-mourns-the-loss-of-professor-peggy-ng/

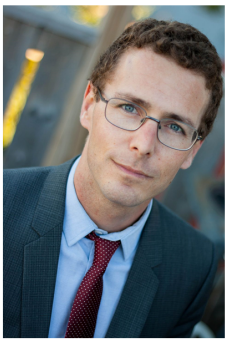
Organizing Committee

Dr. Olli Saarela
(DLSPH, University of Toronto)

Tony Panzarella
(DLSPH, University of Toronto)

Dr. Teresa To
(The Hospital for Sick Children)

Mohammad Kaviul Kahn, PhD candidate
(DLSPH, University of Toronto)



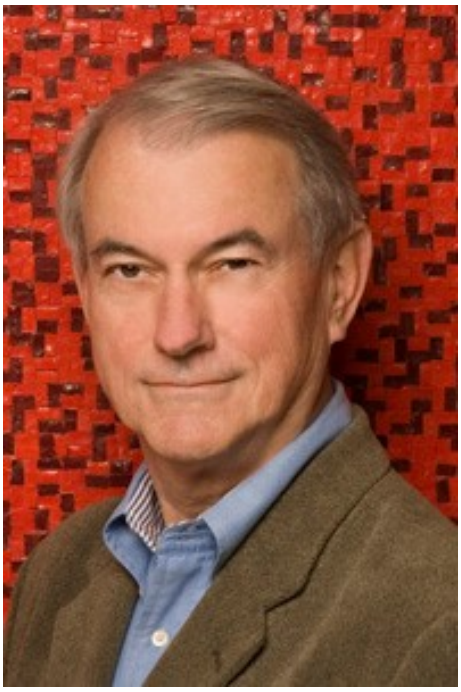
Ryan Rosner
(DLSPH, University of Toronto)

Myrtha Reyna
(The Hospital for Sick Children)

Ashley Mao, MSc candidate
(DLSPH, University of Toronto)

Andrew Tran MSc candidate
(DLSPH, University of Toronto)

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 occupation 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.

Judges

- **Andrew Paterson**
  The Hospital for Sick Children
- **Charles Keown-Stoneman**
  Applied Health Research Centre (AHRC)
- **Erjia Ge**
  Dalla Lana School of Public Health, University of Toronto
- **Katherine Daignault**
  Department of Statistical Sciences at University of Toronto
- **Lisa Strug**
  The Hospital for Sick Children
- **Mohsen Soltanifar**
  University of Toronto & The Hospital for Sick Children
- **Nicholas Mitsakakis**
  Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto
- **Zihang Lu**
  University of Toronto & The Hospital for Sick Children
- **Osvaldo Espin-Garcia**
  Department of Biostatistics, Princess Margaret Cancer Centre, UHN
- **Rahim Moinnedin**
  Department of Family and Community Medicine & Division of Biostatistics, DLSPH
- **Ruth Coxford**
  ICES
- **Sandra Gardner**
  Kunin-Lunenfeld Centre for Applied Research and Evaluation, Rotman Research Institute, Baycrest Health Sciences
- **Shahriar Shams**
  Department of Statistical Sciences, University of Toronto
- **Therese Stukel**
  ICES & Institute of Health Policy, Management and Evaluation, University of Toronto
- **Olli Saarela**
  Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto

Presentation Schedule Session 1

Wednesday, June 17th - 10:00-11:30am EST

No	Title	Presenter	Time
1	Efficacy Assessment in a Clinical Trial with The Use of External Control Arm	Ashley Mao	10:05-10:10
2	A novel clinical classification tool for diagnosing myopathy	Andrew Tran	10:10-10:15
3	Longitudinal Patterns of Distress in Cancer Patients	Jianhui Gao	10:15-10:20
4	Nonlinear Profiles of Cognitive Aging in Healthy Population	Lei Qin	10:20-10:25
5	Distance Concentration and Implications for High-Dimensional Inference	Derek Latremouille	10:25-10:30
6	Artificial neural networks for simultaneously predicting multiple outcomes of symptom burden among patients with cancer	Jennifer Xuyi	10:30-10:35
7	A Bayesian latent class approach to causal inference with longitudinal data	Kuan Liu	15:35-15:40
8	Deriving Dosimetric Predictors of Radiation Induced Toxicity	Jingxian Lan	10:40-10:45
9	Hypothesis Testing in Joint Models for Longitudinal and Time-to-event outcomes	Yuan Bian	10:45-10:50
10	Analytic Challenges On Estimating Effectiveness Of A Regulatory Training Program	Mingxin Sun	10:50-10:55
11	COVID-19 Testing Data and Trends in Canada	Alex Bushby	10:55-11:00
12	Railway Image Segmentation Using Satellite Imagery	Soo Woon Chung	11:00-11:05
13	Longitudinal Growth Clustering in TARGet Kids	Xinyue Chang	11:05-11:10
14	Efficient Parameter Estimation for Individual-Level Models of Infectious Disease Transmission	Madeline Ward	11:10-11:15
15	Deep Tensor Factorization for Survival Prediction of breast cancer using high-throughput omics data	Zhongyuan Zhang	11:15-11:20

Presentation Schedule Session 2

Thursday, June 18th - 10:00-11:30am EST

No	Title	Presenter	Time
1	The cancer screening efficiency of colonoscopy in Lynch syndrome families evaluated by joint models	Yuxin Zhang	10:05-10:10
2	The Interaction of Genetic Risk for Alzheimer's Disease and Glaucoma in Pathological Aging	Amin Kharaghani	10:10-10:15
3	Cost Analysis of Inpatient versus Outpatient Single Level Lumbar Laminectomy Surgery	Yingshi He	10:15-10:20
4	A Simulation Comparison of Conventional and Registry-Based Randomized Controlled Trial Designs for Evaluating Breast Cancer Screening Program	Cheng Wang	10:20-10:25
5	Mendelian randomization to determine a causal relationship between C-reactive protein and diabetic kidney disease	Clarissa Wirianto	10:25-10:30
6	Replacing Experts with Student Raters: an Example Using MRI Carotid Vessel Wall Volumes	Lee Radigan	10:30-10:35
7	Estimating Cause-Specific Mortality Rates by Smoking Levels Using Vital Statistics and . Nationally Representative Survey Data	Linda Luu	10:35-10:40
8	Two-Phase Study Design and Analysis of Quantitative Traits for Multi-region Targeted Genetic Sequencing	Guan Wang	10:40-10:45
9	A Phase II Study of Drug XZ-184 in Recurrent/ Metastatic Endometrial Cancer	Ming Zeng	10:45-10:50
10	Assessing behavioural shift and interference control in children with adhd and the influence of feedback: A pilot study of the flanker task	Saneea Mustafa	10:50-10:55
11	Quantifying the contribution of socioeconomic status to racial blood pressure disparities . in the United States using distributional decomposition	Victoria Tan	10:55-11:00
12	The application of propensity score matching in accessing the efficacy of Fenebrutinib using external control	Yanbo Yang	11:00-11:05
"13	A Multi-modal Assessment of Speed of Processing (SoP) Using Gait, Eye-tracking, and Neuropsychological Measures in a Stroke Cohort	Yuelee(Ben) Khoo	11:05-11:10"
14	Longitudinal Patterns of Distress in Cancer Patients	Siyi Wang	11:10-11:15
15	NASH ICU VS. NON-ICU and Post-transplant Outcomes	Xusiqiao Cai	11:15-11:20

Presentation Schedule Session 3

Thursday, June 18th - 15:00-16:30pm EST

No	Title	Presenter	Time
1	Bayesian tensor factorization-drive breast cancer subtyping by integrating multiomics data	Bowen Cheng	15:05-15:10
2	Genome-wide Association Study of Pseudomonas aeruginosa Infectious in Cystic Fibrosis	Boxi Lin	15:10-15:15
3	Validation of three simulated administrative data algorithms	Di Niu	15:15-15:20
4	Multiple Imputation: Handling Missing Data in Longitudinal Multi-item Scales	Estevam Teixeira	15:20-15:25
5	Cognitive Profiles on Stop Signal Task	Jiachen Zhu	15:25-15:30
6	Prediction of Brain Injury Based on Heart Rate Variability in Hypoxic Ischemic Encephalopathy	Jingwen Du	15:30-15:35
7	Using interactive dashboard to visualize COVID-19 data in Canada	Rose Garrett	10:35-10:40
8	Childhood Vaccination Rates: What Factors May Lead One to Not Vaccinate	Michael Prashad	15:40-15:45
9	Non-Steroidal Anti-Inflammatory Drugs for the Management of Pain Due to Knee and Hip Osteoarthritis: Network Meta-Analysis with Gaussian Random Walk Model	Pai-Shan Cheng	15:45-15:50
10	Two-Stage Joint Modeling of Survival Data and Longitudinal Performance Score for Palliative Care Cancer Patients	Qixuan Li	15:50-15:55
11	Prediction of Spatial Epidemics by a Random Forest Classifier	Salha Qahl	15:55-16:00
12	Development of a R Shiny interactive web app for the Diabetes Population Risk Tool (DPoRT) model	Shuting Lou	16:00-16:05
13	Assessing longitudinal K-Means clustering Clustering Approaches for the CCRS Dementia Dataset	Wenyu Huang	16:05-16:10
14	Financial Burden Among Patients with Renal Cancer in a Publicly Funded Health Care System	Yimin Guan	16:10-16:15
15	Estimate Sociodemographic and Chronic Condition Impact on the Risk of Cardiovascular Disease by Cardiovascular Disease Population Risk Tool	Zhuo Wei	16:15-16:20

Abstracts for Poster Presentations

Abstracts Session 1

Wednesday, June 17th - 10:00-11:30am EST, Chair: Tony Panzarella

Efficacy Assessment in a Clinical Trial with The Use of External Control Arm

Huiyan (Ashley) Mao[1], Melanie Poulin-Costello[2]
1 Dalla Lana School of Public Health, University of Toronto. Toronto, Canada
2 Roche, Canada

Purpose: Randomized controlled trial (RCT) has been the gold standard in the investigation of treatment effect as it is effective to balance out the confounders. Due to the high cost of RCT, more cost-efficient study designs seek methods to pool external dataset. This approach was extended to a Roche study on the indication of Systemic Lupus Erythematosus (SLE), with the use of GSK external control arm. The treatment effect of the investigational drug was assessed with the use of external control arm.
Method: The baseline covariates of patients were balanced using propensity score, which is the probability of being assigned into treatment arm conditioned on observed baseline covariates. Nearest neighbour matching (NNM), NNM with caliper and full matching were applied to form two matched sets: 1) Roche treatment A (200mg BID) and GSK control, and 2) Roche treatment B (150mg QD) and GSK control. The treatment effect was evaluated based on the difference in the response rate between the treatment arm and the external control arm.
Result: The response rate in treatment and control arms was different by 9.76% with odds ratio of 1.48 (95% CI: [0.80, 2.75]) in set 1, and was different by 4.88% with odds ratio of 1.22 (95% CI: [0.66, 2.25]) in set 2.
Conclusion: The study demonstrated the potential of using propensity score matching in a single-armed trial. The treatment effect was not found significantly different between the treatment and the control arm. A higher level of balance needs to be achieved in baseline covariates.

A novel clinical classification tool for diagnosing myopathy using microarray expression profiles

Andrew Tran[1], Christopher Walsh[2], Claudia Dos Santos[2], Pingzhao Hu[1,3]
1 Dalla Lana School of Public Health, University of Toronto, Toronto, Canada
2 Keenan Research Centre for Biomedical Science, St. Michael's Hospital
3 Department of Biochemistry and Medical Genetics, University of Manitoba

Purpose: In the present study, we propose a novel classification tool for predicting muscle disease subtypes using microarray expression data and machine learning algorithms.
Methods: Microarray study data from 42 cohorts was obtained from public repositories. The data contains 1260 unique samples and expression data on 34,099 genes. The data was preprocessed using study-specific batch correction and only 2782 genes were selected for analysis. Data augmentation techniques were used to address class imbalances in the muscle disease subtypes. A support vector machine (SVM) model was trained on two-thirds of the samples based on top genes selected by analysis of variance. The model was validated in the remaining samples using area under the receiver operator curve (AUC).
Results: The AUC ranges from 0.611 to 0.649 in the observed imbalanced data and 0.895 to 0.969 in the augmented data. The SVM model performance improved after being trained on balanced, augmented data compared to the imbalanced, original data. Intensive care unit acquired muscle weakness was the best predicted class (95% AUC CI, 0.999 – 1.000) and congenital muscle disease was the worst (95% AUC CI, 0.907 - 0.955).
Conclusion: This tool addresses an important gap in the literature on myopathies and presents a potentially useful clinical tool for muscle disease subtype diagnosis.

Keywords: muscle disease, machine learning, microarray, clinical tool, biomarker

Longitudinal Patterns of Distress in Cancer Patients

Jianhui Gao[1], Wei Xu[1,2], Madeline Li[2], Osvaldo Espin-Garcia[2]
1 Dalla Lana School of Public Health, University of Toronto. Toronto, Canada.
2 Princess Margaret Cancer Center. Toronto, Canada.

Purpose: Extensive amount of research has shown that people who have intense and long-term stress can have a weakened immune system. Understanding distress patterns of cancer patients is crucial for healthcare teams.
Methods: We use Latent Class Growth Analysis (LCGA) with Zero-inflated Poisson to model depression and anxiety. Implementation is done in Traj, a SAS procedure based on mixture models (Jones, Nagin, and Roeder 2001). Depression and anxiety are analyzed first individually and jointly later given that they are closely related. Number of latent class is selected by Bayes factor $2B_{10}=\vec{\Delta}$(BIC).
Results: Seven latent classes are found in both depression and anxiety. More than half (70.1%) of the patients either have no or very low depression (score <2) throughout their time at Princess Margaret. Similarly, for anxiety, 66.3% of patients have no or low anxiety (score <3). Around 5% and 10% of patients have shown decreasing or increasing trend over time. A small portion of patients exhibit long-term high depression and anxiety (8.4% and 10% respectively). High level of distress (depression and anxiety) is usually associated with low physical flexibility and low income. The posterior correct classification rate is 90% and 85% for depression and anxiety model.
Conclusion: More targeted care is needed for people who indicated low physical ability and low income at the time of the first survey.

Keywords: distress, cancer, longitudinal, anxiety

Nonlinear Profiles of Cognitive Aging in Healthy Population

Lei Qin[1], Sandra Gardner[1,2], Malcolm Binns[1,2]
1 Biostatistics Division, Dalla Lana School of Public Health, University of Toronto, Toronto, Canada
2 Rotman Research Institute, Baycrest Health Sciences, Toronto, Canada

Purpose: In this project we investigate the changes of cognitive functions across multiple domains in aging process with data collected from 180 healthy participants. Hypothetically, cognitive aging processes can follow separate profiles, that are pertinent to different aspects of cognitive functions.
Methods: Visuospatial cognitive data are firstly analyzed with Simple Linear Regression (SLR) with age as predictor in SAS program. Then, Piecewise Linear Regression (PLR) models are developed by introducing paired Breaking Points (BPs) to assess the influence of age intervals on coefficient estimation. Predicted PLR models are ranked by the Root Mean Square Error (RMSE) and presented in contour graphics to discover the age intervals at which models fit the data with best results.
Results: The ROCF scores in immediate and delayed recall have a faster decrease from the early 20s into 30s, different from the long gradual drop of the ROCF copy score. Alternative profiles are detected in analyzing the Group Embedded Figure (GEF) test total Score.
Conclusion: Comparing to SLR models, PLR models show distinct coefficient prediction at certain time intervals, suggesting that the effect of age on visuospatial functions follows nonlinear profiles, especially in visuospatial memories. The uncertainty in PLR models can be explained by the availability and varieties of data, on the other hand, suggests further investigations are needed for covariates analysis.

Keywords: cognitive aging, nonlinear profile, SLR, PLR, RMSE, ROCF, GEF

Distance Concentration and Implications for High-Dimensional Inference

Derek Latremouille[1], Michael D. Escobar[1]
1 Dalla Lana School of Public Health, University of Toronto. Toronto, Canada.

Purpose: Problems in modern science, engineering, and health-care increasingly involve the analysis of data that is high-dimensional in nature. In response, novel methodologies have been proposed for use in these settings. However, it is known that myriad counter-intuitive phenomena can arise in the high-dimensional setting which, if not appropriately accounted for, can compromise both the validity and meaningfulness of conclusions derived. These problems are further exacerbated by the inherent difficulty of both formally and empirically assessing the adequacy of conditions imposed by procedures that are deployed for the analysis of this data. Moreover, it has been argued that the disconnect between the theoretical development of methodology for high-dimensional data and its application in data-analysis has become dangerously large, with the assumptions of these procedures often being subjected to little critical examination.
Methods: While it is often difficult to directly assess assumptions in the high-dimensional setting, we demonstrate that there are numerous types of procedures imposing assumptions that implicitly require the sample data to satisfy stringent distance concentration properties.
Results: We present novel results pertaining to distance concentration, and demonstrate their use in empirically assessing modelling assumptions in areas such as inference for large covariance structure and graphical models. For illustration, two commonplace Microarray datasets are analyzed.
Conclusion: Preliminary analysis suggests that certain models proposed for Microarray data may be misspecified.

Keywords: Distance Concentration, High-Dimensional Inference

Artificial neural networks for simultaneously predicting multiple outcomes of symptom burden among patients with cancer

Jennifer (Wenhui) Xuyi[1], *Rinku Sutradhar* [1,2,3], *Hsien Seow*[2,4]
1 Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Canada.
2 ICES, Toronto.
3 Institute of Health Policy, Management and Evaluation, University of Toronto, Canada.
4 Department of Oncology, McMaster University, Hamilton, ON, Canada.

Purpose: Symptoms burdens can affect treatment and recovery of patients with cancer. Majority of studies model symptom outcomes independently without accounting for possible correlation in symptom burden among measures taken from the same patient. We aim to simultaneously predict symptoms of severe pain, depression and poor well-being among Ontario residents diagnosed with cancer 1 year after their initial diagnosis.
Methods: We built a 3-layer Artificial Neural Network (ANN) model. The input layer consists of the input covariates, the hidden layer consists of 3 nodes, and the output layer consists of 3 nodes each representing a symptom category. We used functions from the R RSNNS package to train the ANN. We split the data into training (n = 35,606) and test sets (n = 10,498). We assessed the model performance with sensitivity, specificity, accuracy, AUC, and calibration.
Results: For each of the pain, depression and well-being outcome, the accuracies were 0.69, 0.65 and 0.64; the sensitivities were 0.61, 0.69 and 0.63; the specificities were 0.69, 0.65 and 0.64; the AUCs were 0.7, 0.7 and 0.74. Patients with lung cancer, late stage cancer, chronic diseases and severe symptom distress during initial diagnosis are more likely to experience the multiple symptoms 1 year later.
Conclusion: The developed ANN model identified cancer patients at high risk of experiencing multiple severe symptom burdens using their clinical information. Further studies should be done to compare the ANN performance with a joint mixed model.

Keywords: Artificial Neural Network, symptom management

A Bayesian latent class approach to causal inference with longitudinal data

Kuan Liu[1,2], Olli Saarela[1], George Tomlinson[1,3], Eleanor Pullenayegum[1,2]
1 Dalla Lana School of Public Health, University of Toronto. Toronto, Canada.
2 Child Health Evaluative Sciences, The Hospital for Sick Children, Toronto, Canada.
3 Department of Medicine, University Health Network. Toronto, Canada.

Purpose: Bayesian causal methods that follow a parametric specification of the joint likelihood of treatment, outcome and covariates, are analytically intractable and often unappealing when face with high-dimensional confounders.
Methods: One possible approach for dimensionality reduction is to model the set of confounders as class indicators in a latent class analysis - interpreting the latent class membership as a confounder. This approach mimics the treatment assignment process often seen in observational studies with administrative data that contain a large number of variables which are indicative of the patient's disease and health status. Treatment assignment under this design follows a clinician-driven decision process, where the treating clinician determines the treatment option based on their classification of the patient's health status using these high-dimensional indicators. In this paper, we consider a causal effect that is confounded by an unobserved, visit specific, latent class in a longitudinal setting.
Results: We formulate the joint likelihood of the treatment, outcome and latent class models conditionally on the class indicators, which permits a full Bayesian estimation of the causal effects. A simulation study is conducted to examine the performance of our proposed method with different levels of confounding and varying numbers of class indicators.

Keywords: Bayesian estimation, Causal inference, Longitudinal data, latent class

Deriving Dosimetric Predictors of Radiation Induced Toxicity

Jingxian Lan[1], Amy Liu[1,2], Olli Saarela[1]
1Dalla Lana School of Public Health, University of Toronto, Toronto, Canada
2Princess Margaret Cancer Centre, Toronto, Canada

Purpose: Developing predictors by using dose-volume data usually poses challenges such as multicollinearity. Some commonly used statistical models provide biased results that affect the accuracy of prediction. In this study, we attempted using functional data analysis for dimension reduction of dose-volume data to overcome this limitation.
Methods: Principal component analysis (PCA), functional principal component analysis (FPCA) and functional partial least squares analysis (FPLS) are used for dimension reduction and developing basis function to represent the dose-volume data. Predicting grade $\geq$ 2 toxicity in the diarrhea for anal canal cancer patients is obtained by logistic regression (LR), principal component logistic regression (PCA-LR), functional principal component - logistic regression (FPCA-LR) and functional partial least squares - logistic regression (FPLS-LR). Predictive performance is assessed by area under receiver operating characteristic curve (AUC), calibration and Brier score.
Results: FPLS-LR demonstrated the significant association between radiation treatment dose and grade $\geq$ 2 toxicity in the diarrhea for anal canal cancer patients. It also provides slightly better predictive performance than LR, PCA-LR, FPCA-LR models based on AUC, calibration and Brier score.
Conclusion: Functional data analysis could overcome the limitations of analyzing the dose-volume effect for toxicity. It can sometimes improve the predictive performance comparing to non-functional analysis.

Keywords: dimension reduction, functional data analysis, functional principal component analysis, functional partial least squares analysis, logistic regression

Hypothesis Testing in Joint Models for Longitudinal and Time-to-event outcomes

Yuan Bian[1,2], Myriam Brossard[2], Shelley Bull[1,2]
1 Dalla Lana School of Public Health, University of Toronto. Toronto, Canada.
2Lunenfeld-Tanenbaum Research Institute, Hospital Mount Sinai. Toronto, Canada.

Purpose: Many clinical studies collect longitudinal biomarkers known to be highly associated with a time-to-event outcome. Current approaches testing for genetic association either analysing longitudinal process and time-to-event process separately or using naïve joint model without taking account of the underlying longitudinal trajectory and interval censoring, leading to inefficient and biased estimates. In this study, we explore different approaches of joint model and using splines in the longitudinal process to develop a robust joint model method for testing genetic association.
Methods: Separate approaches, Two stage approaches, Joint model using Frequentist approaches, and Joint model using Bayesian approaches are implemented during the analysis, along with naive linear time longitudinal model and spline longitudinal model to capture nonlinear subject-specific evolutions in the longitudinal process. Due to the superiority of MCMC approaches, Bayesian approaches can also account for interval censoring.
Result: We proposed and validated a closed form sample size formula for an overall genotype effect in our joint model setting, and included a comprehensive simulation study to evaluate robustness to model misspecification due to non-linearity of the longitudinal traits.
Conclusion: The joint model with splines in the longitudinal process represents a more robust and accurate approaches for testing genetic association.

Keywords: joint model, hypothesis testing, sample size, longitudinal biomarker.

Analytic Challenges On Estimating Effectiveness Of A Regulatory Training Program

Mingxin Sun[1], Victoria Landsman[2], Lynda Robson[2]
1 Dalla Lana School of Public Health, University of Toronto. Toronto, Canada
2 Institute for Work & Health

Purpose: Construction continues to be a hazardous industry these years, the leading cause of traumatic fatalities in construction is a fall from heights, responsible for 38% of incidents. A regulatory training standard for construction workers using fall protection equipment became mandatory in 2015 in the province of Ontario, Canada. In this study, we want to measure the change in safe work practices attributed this training program among learners and explore the determinants of the change based on a longitudinal survey of 633 learners was conducted in 2017 at one-, four-, seven-week post-training and in 2019.
Methods: Factor analysis is used to create continuous measure from 14 items related to safety practices and we use linear mixed-effects models to explore differences in mean scores at different time points.
Results: The participants in this Working at Height training classes changed their practices after they finished the program a few weeks later (p-value$<$0.001 for selected questions). Two years after, these good practices are still maintained.
Conclusion: The present study has provided an example of mandatory training regulations being effective in reducing the incidence of a targeted type of injury. It provides evidence that a mandatory requirement for workers to complete a specified Working at Height training did make a change in safe work practices.

Keywords: Mixed-effects models, factor analysis

COVID-19 Testing Data and Trends in Canada

Alexandra Bushby[1], Kuan Liu[1], Thai-Son Tang[1]
1Dalla Lana School of Public Health, University of Toronto. Toronto, Canada.

Purpose: In the absence of a proven effective vaccine and treatment, contact tracing and testing are the two key measures to control the spread of COVID-19. In this study, we use open-access data to assess testing capacity and geographic differences in testing trends in Canada by provinces and territories to help identify potential gaps in testing strategies.
Methods: Data visualization tools, in conjunction with descriptive data analyses, are used. Such data analyses for each province/territory include, but are not limited to:

- Total number of tests conducted
- Total number of tests conducted per 100,000 population
- Comparison of total tests per 100,000 population to the total number of cases per 100,000
- Total number of daily new tests conducted
- Proportion of patients tested positive out of the number of tests conducted

Results: There are geographic differences in testing capacity across provinces/territories, with Alberta having the highest testing rate per 100,000 population since late March. Significant increase of testing capacity is identified in Quebec and Ontario in early March and late April, respectively. Ontario with the second highest number of confirmed COVID-19 cases consistently ranked last when comparing tests conducted per capita until late April.
Conclusion: Gaps in testing strategies were identified, using data visualization tools and descriptive statistics. These are outlined on our online dashboard.

Keywords: COVID-19, data visualization, testing trends

Railway Image Segmentation Using Satellite Imagery

Soo Woon Chung[1], Jodie Zhu[1], Wendy Lou[1]
1 Dalla Lana School of Public Health, University of Toronto, Toronto, Canada

Purpose: Urban noise has been considered as a serious problem in public health because exposure to loud noise can cause many health problems. CANUE team have developed models that can predict urban noise level. As transportation is the main source of the urban noise, they have used traffic data. To improve the prediction models, we want to utilize satellite imagery to extract meaningful features. In this project, we explore a method of detecting or segmenting railway from satellite imagery.
Methods: Railway image segmentation is performed using U-Net model. It is advanced version of convolution neural network model. First, we prepared training dataset by generating true mask image data using Google Earth Engine. Then, we used FastAI package in Python to train a U-Net model. Due to imbalanced dataset, we used dice score metric.
Results: The dice score was around 0.94. But when we confirmed the results by visualizing three predicted image tiles, railway segmentation was not that great.
Conclusion: We tried hyperparameter tuning but it did not improve our results. To have better prediction model, we may have to use higher resolution imagery or apply data augmentation techniques to overcome class imbalance problem.

Keywords: railway segmentation, satellite imagery, U-Net, convolution neural network, dice score, class imbalance, urban noise, Google Earth Engine

Longitudinal Growth Clustering in TARGet Kids

Xinyue Chang[1], Charles Keown-Stoneman[2]
1 Dalla Lana School of Public Health, University of Toronto, Toronto, Canada
2 Applied Health Research Centre (AHRC), St Michael's Hospital, Toronto, Canada

Purpose: The Applied Research Group for Kids (TARGet Kids!) is a cohort study in examining risk factors related to health equity, routine life exposures and living environment assessment of children. In this research, we firstly measure the availability and feasibility of LCMM approach in longitudinal growth clustering on simulated datasets. Then, this method is applied to real TARGet Kids! research data. The overall objective is to help preschool children improve their health conditions by early identification and preventable interventions through regularly pediatric primary care visits and behavioural counselling.
Methods: Based on the linear mixed model theory, LCMM package in R is primarily used to estimate statistical models along with linear and natural cubic splines. Next, we use the Bayesian information criterion (BIC) as a reference to determine the optimal number of groups. Each subject is classified to the latent class with the highest mean of posterior probabilities calculated according to the Bayes theorem.
Results: The LCMM procedure in clustering analysis remains an over 95% accuracy in the estimation of latent growth curve modelling for generated data. Finally, six sub-groups with distinct patterns over time is explored. High diagonal terms in the posterior classification table indicate a good discrimination.
Conclusion: LCMM functions demonstrate reliable results in growth modelling and developmental trajectory clustering. By linking exposures and growth paths to common health problems, people are benefited from the participation in the study.

Keywords: longitudinal study, LCMM, linear mixed model, trajectory clustering, TARGet Kids!.

Efficient Parameter Estimation for Individual-Level Models of Infectious Disease Transmission.

Madeline A. Ward[1], Lorna E. Deeth[1], Rob Deardon[2,3]
1Department of Mathematics and Statistics, University of Guelph. Guelph, Canada.
2 Department of Mathematics and Statistics, University of Calgary. Calgary, Canada.
3 Department of Production Animal Health, University of Calgary. Calgary, Canada.

Purpose: Individual-level models (ILMs) can be used to model individuals' probability of infection by an infectious disease over time. The ability to account for individual-level information make ILMs very flexible; however, as model complexity and population size increases, likelihood computation times become impracticably long. We propose a new model-fitting method, cluster-disaggregation, and compare its performance to the standard practice of Metropolis-Hastings Markov chain Monte Carlo (MH-MCMC) as well as the increasingly popular Approximate Bayesian Computation (ABC) approach.
Methods: A simulation study is used to compare the performance of the methods to fit a discrete time ILM with a binary susceptibility covariate and a distance-based infection kernel. For the cluster-disaggregation method, data are spatially aggregated into clusters, and the ILM is fit to cluster-level data with MH-MCMC. The cluster-level epidemic curve estimates are then disaggregated to obtain individual-level estimates.
Results: Computation time was decreased considerably by using cluster-disaggregation, however this method tended to produce biased estimates of the total epidemic size and epidemic curve, whereas MH-MCMC and ABC were generally able to produce more accurate estimates.
Conclusion: Each method considered has both advantages and disadvantages that result in no single method being universally recommended over the others.

Keywords: infectious disease modelling, individual-level models, spatiotemporal models, Approximate Bayesian Computation, Metropolis-Hastings Markov chain Monte Carlo

Deep Tensor Factorization for Survival Prediction of breast cancer using high-throughput omics data

Zhongyuan Zhang[1], Wei Xu[1,2], Pingzhao Hu[1,3]
1Dalla Lana School of Public Health, University of Toronto. Toronto, Canada.
2Biostatistics Department, Princess Margaret Cancer Center. Toronto, Canada
3Department of Biochemistry and Medical Genetics, University of Manitoba. Winnipeg, Canada.

Purpose: In cancer survival research, different genomic data have been widely explored, separately. However, genomic data provide complementary individual-specific genomic information, which may include their survival signals. It is challenging to extract valuable information from complex genomic data, which have strong associations with patients' survival outcomes. We proposed an innovative algorithm combining tensor factorization with deep learning, named Deep Tensor Factorization (DTF).
Methods: We built a 3-dimensional (3D) tensor to integrate multi-genomics data and decomposed it into 2D matrices of latent factors that were fed into an autoencoder model to extract hidden features for survival prediction through lasso cox model. We applied the algorithm to the breast cancer data from The Cancer Genome Atlas and evaluated the goodness-of-fit using the concordance index(C-index).
Results: We demonstrated that the proposed tight data integration method shows better prediction performance than other conventional methods. The average C-indexes are 0.655(95% Confidence Interval (CI): 0.616, 0.695) in the DTF model, 0.644(0.598, 0.690) in the TF Model and 0.603(0.533, 0.673) in the Loose integration model.
Conclusion: The DTF model could efficiently integrate different genomic data types for survival analysis and outperforms benchmarked models.

Keywords: breast cancer, omics data, data integration, survival prediction, tensor factorization, deep learning

Abstracts Session 2

Thursday, June 18th 10:00-11:30am EST, Chair: Tony Panzarella

The cancer screening efficiency of colonoscopy in Lynch syndrome families evaluated by joint models

Yuxin Zhang[1], Laurent Briollais[1,2]
1Dalla Lana School of Public Health, University of Toronto. Toronto, Canada.
2 Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, Canada.

Purpose: Colonoscopy is a medical exam to observe abnormalities, such as polyps in the rectum and colon. In this practicum project, we evaluate the screening efficiency of colonoscopy in the members of Lynch syndrome (LS) families, that is to determine whether the screening history of an unaffected member helps predict that member's risk of developing colorectal cancer (CRC). The detection of any polyp and the type of polyp in colonoscopy procedures add up to the screening history of an individual.
Methods: We fit joint frailty models of the recurrent event, which is colonoscopy screening visit after the first visit but before the terminal event or last follow-up, and the terminal event, which is development of CRC. The regression coefficients of frailty models were estimated by maximizing penalized log-likelihood.
Results: We found being male is associated with increased risk of CRC at the 0.05 significant level, and this finding agrees with the results in many other clinical studies.
Conclusion: To conclude, our results suggest that screening history of an unaffected member helps predict that member's risk of developing CRC. Future studies on the screening efficiency of colonoscopy and colorectal cancer could focus on methods to analyze the combination of longitudinal and time-to-event data. Results of the relevant studies could be used to help physicians make clinical decisions and colonoscopy referrals for members of LS families across North America.

Keywords: colorectal cancer screening, lynch syndrome, joint frailty model, recurrent event

The Interaction of Genetic Risk for Alzheimer's Disease and Glaucoma in Pathological Aging

Amin Kharaghani[1], Julie A. Schnieder[2], David A. Bennett[2], Tony Panzarella[1], Daniel Felsky[3]

1 Dalla Lana School of Public Health, University of Toronto, Toronto, Canada

2 Rush University Alzheimer's Disease Center

3 Krembil Centre for Neuroinformatics, The Centre for Addiction and Mental Health

Purpose: Alzheimer's disease (AD) is the most common cause of dementia in aging and has etiological links to primary open-angle glaucoma. Gene variants have been identified as risk factors for both diseases, though the degree of overlap in genetic risk is not well understood. Further, it is not known how risk for these diseases interact to influence pathological aging.
Methods: Using published genome-wide summary statistics, we calculated multiple PRS for AD and glaucoma at multiple p-value thresholds for variant inclusion in two large longitudinal cohort studies of aging including post-mortem amyloid and tau neuropathology. Using linear regression, interaction terms evaluated interdependent effects of these scores on pathological outcomes.
Results: In 2,067 elderly subjects, PRS for AD and glaucoma were correlated (Pearson $r=0.054$, CI95%=[0.011,0.098]) when including a large number of genetic variants in the AD PRS ($p<0.1$) and only the top variants for glaucoma ($p<1x10^{-6}$). A weak interaction was observed between AD and glaucoma PRS (uncorrected $p=0.039$) for levels of tau, whereby individuals at high risk for AD were protected from tau pathology if they had higher risk for glaucoma.
Conclusion: genetic risk scores for AD and glaucoma show weak correlation. Our preliminary results also suggest that risk for one disease may alter the effects of risk for the other on tau-related neuropathology, albeit at uncorrected statistical thresholds.

Keywords: Alzheimer's disease (AD), glaucoma, polygenic risk

Cost Analysis of Inpatient versus Outpatient Single Level Lumbar Laminectomy Surgery

Yingshi He[1,2], Dr. Christopher Witiw[1], Dr. Isaac Aguirre-Carreno[1]
1 Department of Neurosurgery, St. Michael's Hospital, Toronto, Canada
2 Dalla Lana School of Public Health, University of Toronto, Toronto, Canada

Purpose: Expenses in spine surgery are substantially increasing within the last two decades prompting the imperative need for economic evaluations in health care. Our study is to determine whether there is a significant cost difference for inpatient and outpatient treatments for single level lumbar laminectomy from the L1 – S1 (inclusive) levels in a single payer health system where patient-level cost data are available.
Methods: Both parametric and non-parametric bootstrap approaches are applied to compute the difference in mean costs. Regression model is developed to predict patient-level cost.
Results: Non-parametric bootstrap is a better approach for analysis cost data in a small sample size.
Conclusion: Outpatient mean cost is significantly less than inpatient mean cost of single level lumbar laminectomy surgery.

Keywords: single level lumbar laminectomy, cost analysis, inpatient, outpatient.

A Simulation Comparison of Conventional and Registry-Based Randomized Controlled Trial Designs for Evaluating Breast Cancer Screening Program

Cheng Wang[1]
1 Dalla Lana School of Public Health, University of Toronto. Toronto, Canada.

Purpose: Registry-based randomized control trial (RCT) utilizes registry database for trial recruitment, randomization and follow-up. While registry-based RCTs are less expensive and more generalizable compared to conventional RCTs, they rely on registries which are susceptible to bias and incompletion. This study uses simulation to provide quantitative comparisons between conventional and registry-based RCT designs in a hypothetical scenario of evaluating breast cancer screening program.
Methods: An irreversible illness-death model with three states (healthy, disease and death) was used to simulate breast cancer incidence. Time from disease onset to death was the outcome while mortalities from other causes were considered as competing risk. Participants were sampled and followed using both conventional and registry-based RCT approaches. Cost and precision in estimating hazard reduction due to breast cancer screening were compared.
Results: The relative performances of conventional and registry-based RCT designs depended on specific scenarios. Registry-based RCT was cost-effective when bias in the data was low. Conventional RCT provided better performance at higher cost.
Conclusion: Both RCT designs have distinct advantages and disadvantages. Our study suggests that data quality and cost are important factors to consider when deciding which RCT approach to implement. Simulation can provide quantitative evidence to help making such RCT design decision.

Keywords: registry-based clinical trial, simulation, irreversible illness-death model, breast cancer screening

Mendelian randomization to determine a causal relationship between C-reactive protein and diabetic kidney disease

Clarissa Wirianto[1], Delnaz Roshandel[2], Andrew D. Paterson[1,2]
1 Dalla Lana School of Public Health, University of Toronto. Toronto, Canada
2 Genetics & Genome Biology, The Hospital for Sick Children, Toronto, Canada

Purpose: Diabetic kidney disease (DKD) is a common occurrence in people diagnosed with diabetes and causes long lasting negative impacts. Treatment options for end-stage DKD are costly and reduce the quality of life of individuals. Identifying novel approaches for screening and treating DKD is desirable for mitigating the health burden in people with diabetes. This study aims to understand the relationship between the molecular risk factor C-Reactive Protein (CRP) and DKD, which may contribute to developing novel therapeutics.
Methods: Summary statistics from the largest meta-GWAS available for both CRP and DKD were used to perform Mendelian randomization analysis (MR). Random effects inverse variance weighted (IVW) linear regression and Wald ratios were used as methods of MR.
Results: Statistically insignificant regression estimates were obtained from the IVW linear regression. Wald ratio SNPs were also obtained, with rs2064009 out of 43 SNPs showed potential significance.
Conclusion: Lack of significant results suggest no causal relationship between CRP and DKD. Further studies with greater sample sizes and adjusted for covariates such as BMI is required to determine the causal relationship between CRP and DKD.

Keywords: Diabetic kidney disease, C-reactive protein, genome-wide association studies, Mendelian randomization.

Replacing Experts with Student Raters: an Example Using MRI Carotid Vessel Wall Volumes

Lee H. Radigan[1,2], Mariam Afshin PhD[3], Alan R. Moody FRCR[1,3], Pascal N. Tyrrell PhD[1,2]

1 Department of Medical Imaging, University of Toronto, Toronto, Canada.

2 Department of Statistical Sciences, University of Toronto, Toronto, Canada.

3 Department of Physical Sciences, Sunnybrook Health Sciences Centre, Toronto, Canada.

Purpose: In radiology, students are more available and cost effective than radiologists, but are less experienced with larger error. This study aimed to investigate the relationship between number of students and their accuracy when measuring carotid vessel wall volumes in order to explore the potential of replacing an expert.

Methods: A set of 10 students and 1 expert rater recorded measures on a set of carotid artery images. Their ratings were used to estimate the interclass correlation coefficient (ICC), mean absolute error (MAE), and coefficients of variation (CV) between student and expert ratings. The observed CVs were used to estimate student error in order to conduct a simulation to estimate the number of students necessary to obtain similar accuracy as the expert. The simulation was run, and the resulting MAE of the expert was compared to student group MAE as number of students increased.

Results: As number of students increased; ICC increased, and MAE decreased, indicating that averaging student observations tends towards the experts' rating.

Conclusion: Simulation results suggest that the appropriate number of students required to match an expert is dependent on the abilities of the students and the expert.

Keywords: medical imaging, radiology, carotid artery, MRI, intraclass correlation coefficient, coefficient of variation, statistical simulation

Estimating Cause-Specific Mortality Rates by Smoking Levels Using Vital Statistics and Nationally Representative Survey Data.

Linda Luu[1], Victoria Landsman[2]
1 Dalla Lana School of Public Health, University of Toronto. Toronto, Canada.
2 Institute for Work & Health, Toronto, Canada.

Purpose: It is difficult to put the risks associated with cigarette smoking into perspective and fully understand the magnitude of its effects. Know Your Chances is an interactive webpage for customizable risk charts. These charts provide the probability of dying to various causes in a time frame, based on an individuals age, gender and race. Updating this tool to consider prominent risk factors like smoking allows users to better understand the impacts on their health.
Methods: The National Health Interview Survey – Linked mortality file, a nationally representative survey of the United States population linked to the National Death Index, is used to estimate cause-specific hazard ratios (HRs) through Cox regression, and sub-distribution HRs using Fine & Gray methods. The estimates obtained are compared to estimates in published literature that combined several cohorts of volunteers. Mortality rates from census data are split by smoking status using the sub-distribution HRs.
Results: The two datasets produced similar cause-specific HRs for most causes of death, although some estimates from the combined cohort were much higher.
Conclusion: The sub-distribution HRs and cause-specific mortality rates produced in this study can be used to develop absolute risks of dying to various causes for the Know Your Changes webpage.

Keywords: hazard ratios, cause-specific, sub-distribution, mortality rates, survival, cigarette smoking, risk

Two-Phase Study Design and Analysis of Quantitative Traits for Multi-region Targeted Genetic Sequencing

Guan Wang[1], Shelley B. Bull [1,2], Osvaldo Espin-Garcia[3]
1 Dalla Lana School of Public Health, University of Toronto. Toronto, Canada.
2 Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Joseph & Wolf Lebovic Health Complex. Toronto, Canada
3 Department of Biostatistics, Princess Margaret Cancer Center. Toronto, Canada

Purpose: Two-phase sampling is proposed to reduce the aggregated cost of fine-mapping and identify regional genetic associations. We utilize polygenic risk score to extend previous single-region two-phase study designs to multiple genetic regions. The aim of the analysis is to detect genetic regional association of a particular phenotype in large scaled mapping.
Method: The polygenic risk score is calculated by lassosum methods using GWAS summary statistics from Willer and others (2013). To select phase-two data, a residual dependent sampling design is performed by regressing the phenotype on the polygenic risk score and additional covariates. Regional association is carried out under a semi-parametric modelling. The application dataset for fine-mapping analysis is taken from the North Finland Birth Cohort 1966, where attention is paid to triglyceride levels as the phenotype of interest.
Results: The phase 2 sample fraction equals twenty-five percent and fifty percent. To compare the estimation accuracy, the simple random sampling is performed, and a complete analysis is defined as the estimations from whole sequencing data. Given two sample fractions, residual dependent sampling gives more accurate estimation than simple random sampling, because its results are closer to the results from the complete data analysis.
Conclusion: The polygenic risk score dependent sampling can be applied to estimate regional associations in two-phase study. But it requires further analysis. First, we only analyze one phenotype. Second, we only compare one design, residual dependent sampling. There might be other methods to explore. Thus, further analysis of it on this topic.

Keywords: Polygenic risk score, semi-parametric modelling, two-phase study

A Phase II Study of Drug XZ-184 in Recurrent/Metastatic Endometrial Cancer

Ming Zeng[1], Lisa Wang[2]
1 Dalla Lana School of Public Health, University of Toronto. Toronto, Canada
2 Princess Margaret Hospital, Toronto, Canada

Purpose: Endometrial cancer is the highest incidence gynecologic malignancy and remains the fifth most common cancer in women. The relevance of histopathology and molecular characterization in endometrial cancer supports the clinical evaluation in the disease.
Methods: This is a single arm study that evaluated the activity of the multi-targeted inhibitor cabozantinib (XZ-184) in women with endometrial cancer after chemotherapy. The co-primary endpoints were response rate and 12-week progression-free-survival (PFS). Efficacy analysis were conducted on the accrued 102 patients at Princess Margaret Hospital. Patients with rare histology endometrial cancer (including clear cell, carcinosarcoma) will be enrolled in an exploratory cohort.
Results: In serous cohort, 6 patients had responses (cancer shrinks or disappears after treatment). The response rate was 17.6%, 25 out of 34 instances of 12-week PFS were observed and median PFS was 5.2 months. Among 36 patients in endometroid cohort, 6 patients responded at a response rate of 16.7%. 27 patients had 12-week PFS and median PFS was 5.4 months. The waterfall plot indicates that the target lesion size generally decreased dramatically for patients.
Conclusion: Activity can be observed in both serous and endometrioid histology endometrial cancer. The phase II study has a positive result and is worthy of further testing and evaluation in gnomically characterized patient cohorts.

Keywords: endometrial cancer, cabozantinib, response rate, 12-week progression-free-survival

Assessing behavioural shift and interference control in children with ADHD and the influence of feedback: A pilot study of the flanker task

Saneea Mustafa[1], Annie Dupuis[1], Russell Schachar[2]
1 Dalla Lana School of Public Health, University of Toronto. Toronto, Canada
2 The Hospital for Sick Children, Toronto, Canada

Purpose: Attention-Deficit/Hyperactivity Disorder (ADHD) is a common neurodevelopmental disorder impacting children. The onset of ADHD may result in cognitive control deficits and diminished brain activity, impacting their ability to pay attention. In this study, children with and without ADHD were given a cognitive task called the Flanker task to observe and compare their attention and control. Attention was measured through post-error slowing (PES) whereas control was measured through their demonstrated interference control (i.e. ability to ignore distracting stimuli). The study also explored the effect of feedback on participants' PES and interference control.
Methods: During the Flanker Task, a central letter surrounded by distracting letters is presented to a subject. Based on the letter, the subject is asked to press a corresponding button. The outcomes, response time and accuracy, were recorded. Response Time was analysed using Repeated Measures Mixed Model, whereas Response Accuracy was analysed using Repeated Measures Logistic Regression.
Results: Behavioural results from this study showed significant post-error slowing of 50 – 115 ms in both ADHD and control groups. The ADHD group demonstrated less post-error slowing and significantly lower accuracy compared to the control group. The introduction of feedback demonstrated increased post-error slowing and significantly higher accuracy in children with ADHD, whereas feedback only increased accuracy in controls.
Conclusion: Without feedback, both groups demonstrate awareness of mistakes after errors, but the ADHD group demonstrates weaker awareness and lower interference control compared to the control group. Feedback tends to improve awareness of mistakes and interference control.

Keywords: ADHD, Flanker Task, Post-Error Slowing, Interference Control, Response Time, Repeated Measures Mixed Model, Repeated Measures Logistic Regression

Quantifying the contribution of socioeconomic status to racial blood pressure disparities in the United States using distributional decomposition.

Victoria Tan[1], Arjumand Siddiqi[1]
1 Dalla Lana School of Public Health, University of Toronto, Toronto, Canada

Purpose: Black populations in the United States report higher hypertension rates than their White counterparts. Many studies attribute this disparity to the overall lower socioeconomic status (SES) of Blacks, whereas others suggest that socioeconomic advancement for Blacks may lead to higher blood pressures. The uncertainty about SES' impact on blood pressure disparities is due to the lack of a methodology to understand entire blood pressure distributions, rather than hypertension risk or prevalence only. In this study, we present a novel distributional decomposition approach to evaluate the extent that SES accounts for racial blood pressure disparities.
Methods: We decompose racial blood pressure disparities by income, employment, and education status individually and together to create counterfactual blood pressure distributions. Counterfactuals are created by reweighing the socioeconomic profile of Blacks to reflect that of Whites and represent how much racial blood pressure disparities could be mitigated if socioeconomic differences were eliminated.
Results: Socioeconomic differences may better describe racial blood pressure disparities at higher versus lower blood pressure values. Factors other than but potentially related to SES may account more for blood pressure disparities than SES alone.
Conclusion: Black-White racial blood pressure disparities can be explained only in part by socioeconomic differences. A distributional decomposition approach provides nuanced insights into blood pressure disparities that have important implications on improving racial health inequities.

Keywords: Racial health disparities, blood pressure, socioeconomic status, distributional decomposition

The application of propensity score matching in accessing the efficacy of Fenebrutinib using external control

Yanbo Yang[1], Melanie Poulin-Costello[2]
1 Dalla Lana School of Public Health, University of Toronto. Toronto, Canada.
2 PD - Biometrics, Hoffmann-La Roche Limited. Mississauga, Canada

Purpose: With the greater availability of a wide range of data sources, the use of external controls within pharmaceutical development is increasing. The challenge is how to adequately control the potential bias introduced by confounding factors of external control, who are not a part of current randomization. In this study, we explore and study the methods for re-assessing placebo effects across many studies using external controls.
Methods: Propensity score matching (PSM), which allows control of many baseline covariates simultaneously by matching on a single scalar variable, has been widely used to reduce bias when using external control. With the application of PSM, we re-accessed the efficacy of Fenebrutinib in the systemic lupus erythematosus remission using external control from a recently completed clinical trial (NCT01196091).
Results: Results from this study indicate that PSM is possible to address some sources of bias, including calendar time bias, regional bias, and different endpoint bias. However, distinct remaining sources of bias exist after PSM.
Conclusion: The current study looked into completed randomized controlled trials in a meta-analytic fashion using external control data, which reassured that the size of bias with applied mitigation steps in a given scenario could be of acceptable size.

Keywords: randomized controlled trial, external control, logistic regression with regulation, propensity score matching.

A Multi-modal Assessment of Speed of Processing (SoP) Using Gait, Eye-tracking, and Neuropsychological Measures in a Stroke Cohort

Yuelee Khoo[1], Kelly Sunderland[2], Malcolm Binns[1,2], Sandra Gardner[1,2], Ying Chen[3], & ONDRI Team

1 Dalla Lana School of Public Health, University of Toronto, Toronto, Canada.

2 Rotman Research Institute, Baycrest Health Sciences, Toronto, Canada.

3 Queens University, Kingston, Canada.

Purpose: To create SoP behavioural measure composites for clinical use, we investigated shared and unique aspects of SoP underlying neuropsychological, gait, and eye-movement measures in a cohort with mild to moderate ischemic stroke; and examined whether identified SoP aspects were associated with total cerebral White Matter Hyperintensity (WMH).
Methods: This is a cross-sectional, observational study that utilized data obtained via convenience sampling (n = 160). Eight measures were selected from the three modalities. Principal Component Analysis and Canonical Correlation Analysis were conducted on their residuals after adjusting for age, sex, and education. Pearson correlations between the main PCA components and WMH were also examined.
Results: Neuropsychological and gait measures loaded heavily on Component 1 and their canonical variates correlated significantly, potentially reflecting ventral visual stream processing. For Component 2, gait and eye-movement measures loaded in the same direction which could reflect dorsal visual stream processing. For Component 3, gait and eye-movement measures loaded in the opposite direction which could reflect ventral and dorsal attention orienting systems' processing. No significant correlations were observed between the three components and WMH.
Conclusion: behavioural measure composites can be potentially used as a more comprehensive SoP assessment in clinical settings.

Keywords: ischemic stroke, speed of processing, neurodegeneration, multivariate analysis, principal component analysis, canonical correlation analysis

Longitudinal Patterns of Distress in Cancer Patients

Siyi Wang[1], Wei Xu[1,2], Osvaldo Espin-Gracia[2]
1 Dalla Lana School of Public Health, University of Toronto. Toronto, Canada
2 Princess Margaret Cancer Centre, Toronto, Canada.

Purpose: Patient-reported outcomes have been widely used at the Princess Margaret Cancer Center. It is a good way to monitor and manage distress of cancer patients which is important to improve healthcare by generating patient-specific management plan. We use the depression and anxiety scores in the Edmonton Assessment System revised (ESAS-r) tool as surrogate quantitative measures for distress. The objective of the project is to understand patterns of distress and their risk factors.
Methods: We use longitudinal ESAS depression and anxiety scores as primary outcomes. A Two-stage method is employed. In Stage 1, we apply the extended finite mixture model to identify trajectories of distress. In Stage 2, we fit multinomial regression with potential risk factors of distress patterns identified in Stage 1.
Results: We identified a nine-class solution to depression and a six-class solution to anxiety. Most patients have stable distress trajectories over time with few patients maintaining high distress. More patients experience a decreasing rather than an increasing trend in distress. Patients with worse physical symptoms are likely to have higher distress.
Conclusion: Our results from Two-stage method imply that special support should be provided to high distress patients. The majority patients have stable distress trajectories which indicate the importance of baseline distress to predict class membership.

Keywords: distress patterns, cancer patient, finite mixture model, ESAS.

NASH ICU VS. NON-ICU and Post-transplant Outcomes

Xusiqiao Cai[1], Nicholas Mitsakakis[1], Mamatha Bhat[2], Victor Dong[2]
1 Dalla Lana School of Public Health, University of Toronto, Toronto, Canada.
2 Toronto General Hospital Research Institute (TGHRI)

Purpose: Nonalcoholic steatohepatitis (NASH) is a common etiology of liver dysfunction and can result in progression to cirrhosis. The aim of this study is to evaluate the post-liver transplant outcomes of NASH cirrhosis patients admitted to the ICU pre-liver transplant compared with NASH cirrhosis not requiring ICU admission pre-liver transplant.
Methods: We performed a retrospective cohort analysis of all NASH cirrhosis patients who underwent liver transplant within the SRTR databases consisting of three outcomes: (a) 30-day mortality, (b) length of hospitalization post-transplant, and (c) long-term survival. Those variables selected by AIC were entered in an extended Cox regression model with time-dependent variables.
Results: The chi-square test of 30-day mortality (X^2=42.07, p<0.01) and Kruskal-Wallis test (H=751.95, df=2, P< $2.2e^{-16}$) of the length of hospitalization post-transplant were significant in different medical conditions pre-liver transplant. Factors with an increased risk of death included non-ICU hospitalization (HR[time group 1]=1.568, p<0.001, HR[time group 2]=1.263, p=0.042), and pre-transplantation ICU (HR[time group 1]=1.991, p<0.001, HR[time group 2]=1.399, p=0.029).
Conclusion: 30-day mortality and the length of hospitalization post-transplant was significantly different among the three medical condition groups. Patients transplanted from the ICU and Hospitalized (not ICU) portends a significantly higher risk of death in the first 6.3 years after transplantation.

Keywords: SRTR, NASH, Survival, Post-transplant outcomes

Abstracts Session 3

Thursday, June 18th - 15:00-16:30pm EST, Chair: Tony Panzarella

Bayesian tensor factorization-drive breast cancer subtyping by integrating multiomics data

Bowen Cheng[1], Pingzhao Hu[1,2]
1 Dalla Lana School of Public Health, University of Toronto, Toronto, Canada
2 Department of Biochemistry and Medical Genetics, University of Manitoba

Purpose: Breast cancer is a highly heterogeneous disease. Subtyping the disease and identifying the genomic features driving these disease subtypes are the key for precision oncology for breast cancer.
Methods: We proposed to use Bayesian tensor factorization (BTF) to integrate multiomics data of breast cancer, which include expression profiles of RNA-seq, copy number alteration and DNA methylation measured on 771 breast cancer patients from The Cancer Genome Atlas. We used a consensus clustering approach to identify breast cancer subtypes using the factorized latent features. Subtype-specific survival patterns of the breast cancer patients were evaluated using Kaplan-Meier (KM) estimators. We will use gene set enrichment analysis to identify breast cancer subtype-specific gene pathways. The proposed approach will be compared with other state-of-the-art approaches for cancer subtyping.
Results: The BTF analysis identified 51 optimized latent components, which were used to reveal five major breast cancer subtypes. KM analysis of the cancer subtypes showed distinct survival patterns among the cancer subtypes ($p<0.05$). Differential analysis identified common tumor-suppressing genes (e.g. TSSC1) across the subtypes and subtype-specific cancer genes.
Conclusion: Our preliminary data showed that the proposed approach is a promising strategy to efficiently use publicly available multiomics data to identify breast and other cancer subtypes.

Keywords: Breast cancer subtyping, Bayesian tensor factorization, consensus clustering, survival analysis, multiomics data, gene pathways

Genome-wide Association Study of Pseudomonas aeruginosa Infectious in Cystic Fibrosis

Boxi Lin[1], Lei Sun[1,2], Lisa Strug[1,3]
1 Dalla Lana School of Public Health, University of Toronto, Toronto, Canada.
2 Department of Statistical Sciences, University of Toronto, Toronto, Canada.
3 Genetics & Genome Biology Program, The Hospital for Sick Children, Toronto, Canada.

Purpose: Age at acquisition of Pseudomonas aeruginosa (Pa) infection is an important trait of cystic fibrosis (CF) due to its strong association with severity of CF lung disease later in life. Previous study has shown that genetic modifiers totally play a significant role in the timing of Pa infection in CF individuals, yet optimal, individualized treatment of CF will require identification and targeting of disease modifiers.
Methods: We conduct genome-wide association study (GWAS) with CF patients from Canada in search of genetic modifiers of Pa age, based on linear mixed effect model that accounts for sample relatedness and population structures.
Results: We identify one novel genome-wide significant SNP on Chromosome 21. We also estimate the narrow sense heritability of Pa is as high as 70%. Conclusion: Through the GWAS on Pa, we identify the significant region which deserves further biological examination. And we verify the contribution of genetic modifiers for Pa, and find evidence of polygenic effects.

Keywords: Pseudomonas aeruginosa, GWAS, heritability, cystic fibrosis.

Validation of three simulated administrative data algorithms

Di Niu[1], Rahim Moineddin[2]
1 Dalla Lana School of Public Health, University of Toronto, Toronto, Canada
2 Department of Family & Community Medicine, University of Toronto, Toronto, Canada

Purpose: Administrative data algorithm finds patients with certain conditions in the administrative databases like government insurance plan databases. This study validated three simulated administrative data algorithms against a gold standard.
Methods: The validation process consists of unpaired design and paired design. Unpaired design assumes each algorithm is tested on a different sample of individuals. Paired design assumes all three algorithms are tested on one sample. Diagnostic accuracy measures such as sensitivity and specificity are found under unpaired design. Paired design includes classical tests and logistic regression. Classical tests consist of McNemar test and likelihood ratio test. Logistic regression includes individual hypothesis tests and global hypothesis tests.
Results: For unpaired design, algorithm 2 had highest specificity of 86.4% and algorithm 3 had highest sensitivity of 94.4%. For paired design, individual hypothesis tests showed sensitivity or specificity of all three algorithms were different from one another except the specificity of algorithm 2 and 3. Global hypothesis tests agreed with individual hypothesis tests except that all three specificities were the same.
Conclusions: It largely depends on the purposes of testing to determine which algorithm to use. Logistic regression is more flexible than classical tests as it can test more than two sensitivities or specificities simultaneously. Policymakers and researchers will find these findings useful to study trends of disease outcomes.

Keywords: administrative data algorithms, diagnostic accuracy measures, McNemar test, likelihood ratio test

Multiple Imputation: Handling Missing Data in Longitudinal Multi-item Scales

Estevam C. M. Teixeira [1], Rosane Nisenbaum [1, 2]
1 The Dalla Lana School of Public Health, University of Toronto. Toronto, Canada.
2 MAP Centre for Urban Health Solutions, St. Michael's Hospital. Toronto, Canada.

Purpose: Missing data are a problem that affects almost all real datasets and healthcare data is not an exception. Missingness is a problem that needs to be addressed correctly. Most predictive methods require a complete data frame of predictors and when the proportion of missingness is large, filtering can remove a large proportion of data leading to a decrease in the predictive power of the final model.
Methods: Using a pre-post cohort study from the Coordinated Access for the Homeless (CATCH) program, we investigated the performance of the multilevel multiple imputations to see whether there has been any improvement in people's perception of their quality of life after joining the CATCH program using linear mixed-effect models.
Results: We have identified and compared 3 different multiple imputation approaches for imputing derived variables in longitudinal studies. Analysis of the observed data suggested an improvement in people's perception of their quality of life after joining the program.
Conclusion: Ignoring missing data may cause bias of unknown size and direction in longitudinal studies. The distinction between missing data mechanisms is important to understand why some methods work, and others do not, and whether a missing data method can provide valid statistical inferences.

Keywords: missing data, multilevel multiple imputations, derived variables, longitudinal data.

Cognitive Profiles on Stop Signal Task

Jiachen Zhu [1], Annie Dupuis [1], Russell Schachar [2]
1 University of Toronto, Toronto, Canada.
2 The Hospital for Sick Children, Toronto, Canada.

Purpose: Diagnosis using a subjective measure in psychological disorders like Autism Spectrum Disorder (ASD), Attention-Deficit Hyperactivity Disorder (ADHD) and Obsessive-Compulsive Disorder (OCD) is the conventional method. However, it takes a relatively long time to make such conclusions on diagnosis, and there are minimal objective evidences to supplement that decision. Measures from Stop Signal Task (SST) provides a possibility of obtaining an objective measure.
Methods: First, we treated the data obtained from SST by factor analysis and then applied hierarchical cluster analysis to separate different patient clusters. Finally, we used pairwise comparison technique on the clusters to identify the causes of differences in probability of diagnosis.
Results: Young children who performed poorly in the SST would be more likely to be diagnosed with ADHD.
Conclusion: From the calculated probabilities, age, gender and variable differences, we conclude that relative to the baseline, children with poor performances in the SST variables, especially in stop signal response time (ITSSRT) would lead to a higher chance of being diagnosed with ADHD. The highest probabilities of diagnosis are usually accompanied by both young age and high values in the SST variables.

Keywords: neurodevelopmental disorder, psychological disorder, ADHD, OCD, ASD, hierarchical cluster analysis, factor analysis.

Prediction of Brain Injury Based on Heart Rate Variability in Hypoxic Ischemic Encephalopathy

Jingwen Du[1,2], Nicholas Mitsakakis[2], Ipsita Roy Goswami[1], Emily Tam[1]
1 Neurology Program, The Hospital for Sick Children, Toronto, Canada.
2 Dalla Lana School of Public Health, University of Toronto. Toronto, Canada.

Purpose: Hypoxic ischemic encephalopathy (HIE) is the major cause of death in neonates (an infant less than four weeks old). This study aimed to select significant heart rate variability (HRV) features and clinical factors in neonates with HIE to predict brain injury patterns.
Methods: HRV features were recorded at seven time points (12, 18, 24, 30, 36, 42, and 48 hours of life). Brain injury patterns diagnosed by MRI were classified as normal, focal infarct, and abnormal. Multinomial logistic regression using lasso regression were performed to select HRV features and clinical factors.
Results: 103 neonates were analyzed (109 neonates and 6 were missing). The models predicted well (difference between validation and test error < 0.13) at seven time points except at 12 hours' time point. The normal brain injury patterns were overpredicted (predicted number $>$ actual number); the focal infarct and abnormal brain injury patterns were underpredicted (predicted number $<$ actual number).
Conclusion: Clinicians can record significant HRV features and clinical factors of neonates at seven time points except at 12 hours and use these predictors to predict the brain injury patterns.

Keywords: neonatal, hypoxic ischemic encephalopathy, heart rate variability, brain injury, MRI, feature selection, lasso regression, multinomial logistic regression.

Using interactive dashboard to visualize COVID-19 data in Canada

Rose Garret[1,2], Kuan Liu[1,2]
1 Dalla Lana School of Public Health, University of Toronto. Toronto, Canada.
2 Child Health Evaluative Sciences, The Hospital for Sick Children, Toronto, Canada.

Objective: We are constantly being inundated with COVID-19 news updates, and it can be difficult to make sense of all this information. For this project, we built an interactive dashboard to summarize and visualize COVID-19 pandemic data trends across Canada, with more detailed reporting on highly affected regions.
Methods: We used open-access data provided by the COVID-19 Canada Open Data Working Group in our study. We created an interactive data dashboard application using Rstudio Shiny, Plotly and Leaflet data visualization packages. Our web application was deployed for public viewing with daily data updates.
Results: The interactive dashboard is hosted at https://kuan-liu.shinyapps.io/*canada_dash*/. We provided visualization at national level, at provincial level for Ontario and Quebec and at city level for Toronto, on the cumulative distribution of confirmed cases overtime, daily new confirmed, recovered and deceased cases as well as the daily percentage change on confirmed and deceased cases. In addition, we provided a geographic case map, case trajectory comparison with selected countries, and heatmaps visualizing confirmed and deceased case numbers by health regions for Ontario and Quebec.
Conclusion: We used interactive data visualization to summarize COVID-19 data in Canada. The interactive feature of our visualization will assist general public to efficiently and effectively grasp the current status and trend of COVID-19 in Canada.

Keywords: COVID-19, Data visualization, Shiny Application

Childhood Vaccination Rates: What Factors May Lead One to Not Vaccinate

Michael Prashad[1], Gerald Lebovic[2]
1 Dalla Lana School of Public Health, University of Toronto. Toronto, Canada.
2 Applied Health Research Centre, St. Michael's Hospital, Toronto, Canada.

Purpose: Vaccination rates among young children in USA have been declining in recent years. This study aims to build a predictive model to predict whether a child will receive all recommended vaccinations, only some recommended vaccinations, and flu vaccinations, and to identify factors associated with children not receiving these vaccinations.
Methods: Using survey data from 2012-2017 on children in USA between 19 and 35 months of age, survey-weighted logistic regression models are used to predict a child's vaccination status, and to identify factors which are related to childrens' vaccination status.
Results: The predictive models had poor discrimination, with AUC values around 0.629. Important factors associated with recommended childhood vaccination include family income, number of children in the household, child's age, whether the child was firstborn ($p = 0.032$), mother's age and education, state of residence and whether the child has moved state ($p < 0.001$ unless otherwise stated). The same applies for flu vaccinations, except race and ethnicity were additional factors.
Conclusion: Survey-weighted logistic regression is not a viable method to predict vaccinations in children. Furthermore, attention should be given to areas with poor income, low education, younger children, and highly populated households, to increase vaccination rates among children least likely to get vaccinated.

Keywords: childhood vaccination, predictive modelling, logistic regression, survey weights.

Non-Steroidal Anti-Inflammatory Drugs for the Management of Pain Due to Knee and Hip Osteoarthritis: Network Meta-Analysis with Gaussian Random Walk Model

Pai-Shan Cheng[1], George Tomlinson[1], Bruno R da Costa[2]
1 Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto. Toronto, Canada.
2 Applied Health Research Centre, Li Ka Shing Knowledge Institute, St. Michael's Hospital. Toronto, Canada.

Purpose: Non-steroidal anti-inflammatory drugs (NSAIDs) are commonly prescribed to help patients with knee and hip osteoarthritis manage pain. To assess the effectiveness of different preparations of NSAIDs, a network meta-analysis (NMA) using Gaussian random walk model was conducted.
Methods: The primary outcome data on pain intensity was extracted at up to seven time points after start of treatment from 76 large randomized trials. A Gaussian random walk NMA model was developed and applied to the 76 trials comparing several doses of various NSAIDs, paracetamol, and placebo. The proposed model is an improved formulation over another random walk model previously used to analyze these same trials (da Costa et al., 2017).
Results: The relative treatment effects of various interventions compared to placebo were estimated and interventions were ranked according to their median rank. Top three interventions are etoricoxib 90 mg, rofecoxib 50 mg, and diclofenac 150 mg; while placebo, naproxen 750 mg, and paracetamol <2000 mg rank at the bottom.
Conclusion: Results of this NMA add to existing evidence of the effectiveness of NSAIDs, though clinicians should still consider available safety data in deciding which intervention to prescribe.

Keywords: mixed treatment comparison, random walk, OA.

Two-Stage Joint Modeling of Survival Data and Longitudinal Performance Score for Palliative Care Cancer Patients

Qixuan Li[1], Wei Xu[1,2], Lisa W. Le[2]
1 Dalla Lana School of Public Health, University of Toronto, Toronto, Canada.
2 Princess Margaret Cancer Centre, Toronto, Canada.

Purpose: Several recent studies have explored the relationship between baseline measures of performance status with the survival of patients with advanced cancer. However, the longitudinal progression of performance status is also associated with the risk of death. In this study, we implement a two-stage joint model to explore the relationship between the longitudinal progression of the performance status and the overall survival of palliative care cancer patients.
Methods: A two-stage joint model is established to analyze the longitudinal measures of Palliative Performance Scale(PPS) and patient overall survival simultaneously. The stage one linear mixed model and the stage two cox proportional hazards model are linked by the aggregated fixed time effect and the rate of change of time on PPS score. The main parameter of interest is the effect of the longitudinal change in PPS score on overall survival of patient.
Results: There is significant association between the survival of patients and the change in PPS scores. The overall PPS score has a decreasing pattern over time. The higher rate of change in PPS score tend to decrease the hazard of patients.
Conclusion: The longitudinal progression of PPS score is associated with the overall survival of palliative care cancer patients.

Keywords: palliative care, joint modelling, performance status, survival.

Prediction of Spatial Epidemics by a Random Forest Classifier

Salha Qahl[1], Rob Deardon[1]
1 Department of Mathematics & Statistics, University of Calgary

Mathematical modelling of infectious disease transmissions in public health decision-making has a long history, but has become more common during the past three decades. In real populations however, the true underlying epidemiological process is unknown. Rather, the underlying process must be inferred from the data. One way to address this is to examine how well epidemic models of a known structure can be predicted using statistical classification algorithms. Individual Level Models (ILM's) are stochastic spatio-temporal models that can capture space-time dependence of an infectious disease, and yield insight into how a disease can progress. To examine the effect of population spatial structure, we simulated individuals over a grid and in mixture of multivariate normal clusters. Sample spatial structure was examined by comparing spatial clusters, circular and rectangular stratification at a range of scales. We manipulated model parameter combinations and complexity of the training data set to examine how well Random Forests could accurately classify models under different population and spatial sample structure. Random Forests could consistently identify models with the same transmission distance, and with the same error variance. However, Random Forest failed to discriminate between models with different infectious periods.

Keywords Supervised learning, Spatial Epidemic, transmission model, Random forests, Spatial Stratification.

Development of a R Shiny interactive web app for the Diabetes Population Risk Tool (DPoRT) model

Shuting Lou[1], Brendan Smith[2], Lennon Li[3]
1 Dalla Lana School of Public Health, University of Toronto. Toronto, Canada.
2 Health Promotion, Chronic Disease and Injury Prevention, Public Health Ontario, Toronto, Canada.
3 Dalla Lana School of Public Health, University of Toronto. Toronto, Canada.

Purpose: Reducing unjust difference in health is an essential issue to help people maintain better health status. Reducing health inequities involves two actions: 1) describing a difference in health and 2) identifying which differences in health are considered unjust.
Methods: The estimated individual level 10-year diabetes risks are calculated based on Diabetes Prediction Risk Tool (DPoRT) model and 2015-2016 CCHS data. We applied targeted weight loss percentage intervention with sufficiency and simple equality ethic criteria and compare the level of difference in the risk of diabetes among different risk groups.
Results: A R Shiny web app is developed based on researchers' needs and acts as an interactive tool for researchers to visualize the results graphically based on their needs. The app provides two plots of diabetes risk vs. percentage weight loss for two different ethic criteria.
Conclusion: The graphical representation of the weight loss intervention on the risk for diabetes provides a more straightforward way to understand the effects of the interventions on different target groups.

Keywords: survival analysis, web app, CCHS survey, diabetes risk prediction, health equity

Assessing longitudinal K-Means clustering: Clustering Approaches for the CCRS Dementia Dataset

Wenyu Huang[1], Rafal Kustra[2]
1 Canadian Institute for Health Information, Canada.
2 Dalla Lana School of Public Health, University of Toronto. Toronto, Canada.

Purpose: The treatment of dementia is an increasingly severe problem that was placing a substantial public health threat in Ontario, Canada and the international health care system. Who regards to Alzheimer's treatment as a public health priority, and as the population ages, it is expected that by 2030, there will be more than 70 million Alzheimer's patients in the world. In this study, we will describe and sort out the data of dementia's patients provided by CCRS (The Canadian Care Reporting Systems) throughout Canada and perform an effective classification method to classify the curves of dementia's patient. We will compare the results of the classification to find out the trajectories of the cognitive impairment and the aggressive behaviour caused by the dementia issue through the age in distinct clusters.
Methods: Our primary method would be partition-based clustering, mostly non-parametric, such as K-Means can be applied to our dataset with longitudinal clustering algorithms to _nd the groups that can maximize the similarities within each cluster. K-means algorithm is a mountain climbing algorithm following the principle of expectation maximization Results: around 90 percent of curves can be explained by upwarding CPS scale trend, and 80 percent of patients have up and down ABS scale trend.
Conclusion: From the K-Means Longitudinal clustering results and previous discussions. We can see that, on average, cognitive impairment in dementia gets worse with age. But most people do not have a serious aggressive response, and only a small number of people have some aggressive response in the early stages of the disease, but they all show a downward trend.

Keywords: longitudinal K-Means. Dementia issue

Financial Burden Among Patients with Renal Cancer in a Publicly Funded Health Care System

Yimin Guan[1], Dr. Wei Xu[1,2], Lisa W. Le[2]
1 Dalla Lana School of Public Health, University of Toronto. Toronto, Canada
2 Princess Margaret Cancer Center, Toronto, Canada.

Purpose: The huge rise of treatment costs affects the risk of financial toxicity and patients face severe financial distress. In this study, we aimed to seek the factors associated with greater financial distress in renal cancer patients in the context of Canadian public health care system.
Methods: Financial burden was measured by Comprehensive Score for Financial Toxicity. Demographic variables were analyzed using descriptive statistics and frequency tabulation. Logistic regression models were implemented to determine the variables that associated with greater financial distress (median COST score < 20.5).
Results: The median of age was 64 years, 80.4% were male. In univariable analysis, financial burden was associated with age, total OOP costs and income source. Patients younger than 64 years, paid higher OOP costs and got unemployed suffered worse financial well-being. In multivariable logistic regression analyses, younger age was the only significant factor associated with greater financial distress
Conclusion: Patients with renal cancer younger than 64 reported greater financial burden. Higher OOP costs and unemployed status had a trend towards an association with higher financial burden.

Keywords: renal cancer, financial burden, Comprehensive Score for Financial Toxicity, logistic regression

Estimate Sociodemographic and Chronic Condition Impact on the Risk of Cardiovascular Disease by Cardiovascular Disease Population Risk Tool

Zhuo Wei[1], Brendan Smith[2], Lennon Li[2]
1 Dalla Lana School of Public Health, University of Toronto, Toronto, Canada.
2 Public Health Ontario, Toronto, Canada

Purpose: The Cardiovascular Disease Population Risk Tool (CVDPoRT) is a validated predictive model to predict CVD event in 5 years for population implement. Studies suggest low education, obese, diabetes and hypertension are highly associated with CVD events. I aim to demonstrate those hypotheses by the CVDPoRT algorithm.
Methods: The preliminary preparation is to test my R code on a dummy dataset. The primary analysis is to compute education, BMI, diabetes and hypertension's impact on the CVD risk in 5 years by the CVDPoRT on the Canadian Community Health Survey (CCHS) 2013-2014 dataset.
Results: Based on calculation of multiple predictors' impact on the CVD risk in 5 years, this project provides additional evidence that low education, overweight, diabetes, hypertension are associated with higher CVD risk.
Conclusion: This project demonstrates that predictors like low education, overweight, diabetes and hypertension, are positively associated with the CVD risk. Moreover, the R-shiny web-based application allows users to obtain the risk and expected number results interactively and graphically by selecting predictors. Further research is strongly encouraged due to current limitations. Overall, the combination of R shiny and complex risk-prediction algorithm on population-level datasets have potentials to provide public education and individual health prediction.

Keywords: BMI; Cardiovascular Disease; Diabetes; Education; Hypertension; Survival Analysis

***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