CANSSI Distinguished Postdoctoral Fellows Projects List

Project Name: Development of innovative Bayesian methods to address statistical challenges in longitudinal electronic health record data

Supervisor:

- Name: Zihang Lu
- Affiliation: Assistant Professor, Department of Public Health Sciences & Department of Mathematics and Statistics, Queen's University
- Email: zihang.lu@queensu.ca
- Website: https://phs.queensu.ca/faculty-research/zihang-lu

Co-supervisor:

- Name: Kuan Liu
- Affiliation: Assistant Professor, Dalla Lana School of Public Health & Institute of Health Policy, Management & Evaluation, University of Toronto
- Email: kuan.liu@utoronto.ca
- Website: https://www.dlsph.utoronto.ca/faculty-profile/liu-kuan/

Location/University:

- Queen's University
- University of Toronto

Project abstract

The use of Electronic Health Records (EHR) has become increasingly common in medical research and it is viewed as an important data source for scientific discovery. Several Bayesian methods for analyzing EHR data have been proposed recently. Despite these advances, there remain several methodological research gaps especially for longitudinal EHR data and poor uptake of Bayesian methods for analyzing EHR data in practice, due to hurdles such as generalizability, computational challenges, interpretability, and the lack of user-friendly statistical software.

To this end, the postdoctoral fellow will work on developing innovative, flexible, generalizable and interpretable Bayesian methods to analyze large and complex real-world clinical and public health data. The methodological contribution will focus on patient phenotyping and causal inference under longitudinal settings. This research also includes the development of accompanying software packages, open-access code, and tutorials to facilitate a wide application of the proposed methods.

Plan for interdisciplinary/applied experience

This proposal facilitates new collaboration among faculty members and PDF from multiple institutes with appointments across several departments including statistics, biostatistics, and

medicine. Apart from working with the supervisory team (Dr.s Lu, Liu and Stephens), the PDF will work collaboratively with Dr. Eddy Fan (University of Toronto) and Dr. Laveena Munshi (University of Toronto) and Dr. Geoffrey Anderson (University of Toronto). Drs. Fan and Munshi are leading researchers in critical care medicine who will facilitate access to intensive care EHR data and provide clinical expertise on the application of the proposed research. Dr. Anderson is a Professor in Health Services Research and the lead PI of a CIHR-funded project studying the trajectory of cognitive and functional declines using the Canadian Longitudinal Study of Aging data.

The proposed research will be carried out at both Queen's University and the University of Toronto. Drs. Lu and Liu will be the supervisor and co-supervisor who will lead the training and work closely with the PDF on a daily basis. Dr. Stephens will serve as a research mentor and project collaborator. He will provide methodological guidance and professional development support including career development on a monthly basis. During the analysis stage, monthly or bi-monthly virtual team meetings with clinical and public health collaborators will also be organized.

This interdisciplinary training partnership will foster an interdisciplinary mentorship environment and promote training experience to equip the fellow with both technical and professional skills to succeed in academia and industry as an independent scientist. Training will encompass further development of technical and scientific communications skills (including non-technical presentation, grant and proposal writing), teaching and mentorship skills, leadership and collaborative research skills. There will be ample opportunities for the PDF to be involved in other collaborative biostatistical research projects at the host universities. The PDF will also be invited to participate in grant applications to gain grantsmanship and establish long-term relationships with project collaborators.

Plan for teaching/training/education

The planned training will take place at Queen's University (Kingston, ON) between September and April in Year 1 and Year 2 and at the University of Toronto (UofT , Toronto, ON) between May and August in Year 1 and Year 2. The PDF will be engaged in a variety of teaching and mentorship activities. Specifically, the PDF will teach one three-credit course in Biostatistics each year at Queen's University. In addition, together with Drs. Lu and Liu, the PDF will participate in organizing and teaching a one-day pre–conference short course on applied Bayesian statistics in Year 2.

The PDF will be provided with opportunities to engage in primary supervision of summer undergraduate research students in statistics at UofT, and joint supervision of graduate students in biostatistics from supervisors' research groups at Queen's University and UofT throughout the two-year training program.

The PDF will be funded to attend and present at national and international conferences every year, such as the SSC Annual Meeting, the Joint Statistical Meeting, and the Eastern North American Region Biometric Society Meeting. Additionally, PDF will be encouraged and supported to present and participate in various research seminars and workshops within and outside the host institutes, such as Biostatistics research seminars, clinical epidemiology rounds, and the CANSSI research day.

Plan for mentoring
The PDF will be supported to direct their own research, engage in problem solving and critical scientific appraisal of the work. All technical training uses a structured open-door policy, with team meetings between PDF and supervisors scheduled once a week and additional support offered by email or in-person as needed. Monthly meetings between PDF, supervisors and mentors will support the ongoing research and professional training progress.

Every year the PDF will formally discuss career goals and how training can be enhanced to meet specific goals. Each term the PDF will be supported to participate in 1-2 professional development courses/workshops offered by the School of Graduate Studies at the two host universities. This will include courses on Teaching in Higher Education (Year 1) and the Equity, Diversity, and Inclusion training workshops (Year 1 and Year 2).

Drs. Lu and Liu will function as supervisor and co-supervisor who will lead the PDF training and work closely with the PDF on a daily basis. Dr. Stephens will function as a research mentor and collaborator who will provide methodological guidance and professional development support including career development on a monthly basis. During the analysis stage, monthly virtual team meetings with clinical collaborators (Drs. Fan and Munshi) will also be organized.

Queen's University and UofT are Canada's leading research-intensive universities with a global reputation for cutting-edge collaborative research intertwining statistics, data sciences, medicine and public health. The PDF will have access to seminars, training, scholarships, networking opportunities, collaborations, and peer support through various initiatives across the two universities including the Data Sciences Institute and Vector Institute at Toronto and the Canadian Cancer Trials Group at Queen's University.

Proposed schedule
Schedule for Year 1: **(a) Research**: The PDF will work on developing a novel Bayesian unsupervised learning methods for electronic health record data with complex structures (Objective 1). Under the supervision of Drs Lu, Liu and Stephens, the PDF will work on the model specification, developing efficient algorithms and software code in Year 1 (Sep-Apr). This newly developed method will be applied to real data applications (Year 2, May-Aug). **(b) Teaching**: The PDF will teach a full course (three credits) at Queen's University and engage in supervising graduate students on their research projects. **(c) Other activities**: Prepare the manuscript and submit it for publication. Attend and present research findings in seminars, workshops and conferences. Help organize seminars and workshops.

Schedule for Year 2: **(a) Research**: The PDF will work on developing novel Bayesian methods for causal inference (Objective 2). Under the supervision of Drs Lu, Liu and Stephens, the PDF will work on the model specification, developing efficient algorithms and software code in Year 2 (Sep-Apr). This newly developed method will be applied to real data applications (Year 2, May-Aug). **(b) Teaching**: The PDF will teach a full course (three credits) at Queen's University and engage in supervising graduate students on their research projects. **(c) Other activities**: Prepare the manuscript and submit it for publication. Attend and present research findings in seminars, workshops and conferences. Help organize seminars, workshops, and pre-conference short courses.

List of qualifications of suitable candidates

- A doctoral degree in Biostatistics or Statistics
- Strong programming skill; Proficient in R, C++ and SAS
- Knowledge of Bayesian statistical methods and modelling
- Strong communication, written and analytical skills
- Experience in collaborative research
- Ability to independently organize workload, set goals and work effectively towards deadlines
- Experience with analyzing clinical and public health data is an asset

Research description

Background: Electronic Health Record (EHR) data has become increasingly common and it has been viewed as an important data source for scientific discovery in medical research. EHR is a database that collects information such as demographics, diagnostic billing codes, medication, procedure codes, vital signs, laboratory test results, and clinical imaging. These data are longitudinal by nature, providing real- world evidence on disease development, progression and response to interventions. EHR collection is primarily for supporting the care of patients rather than for scientific research on treatment or intervention development. Consequently, issues such as selection bias, confounding bias, information bias, misclas- sification, and missing data hinder the direct application of existing statistical methods [1]. Moreover, large sample sizes, repeated measures, irregular visit time, and unbalanced number of observations pose additional modeling and computational challenges when analyzing these data to inform clinical decisions. Recently, there are several studies developing Bayesian methods for analyzing EHR [2, 3, 4, 5]. *Despite this advancement, Bayesian methods still receive very limited attention and have not been widely used in practice due to hurdles such as generalizability, computational challenges, interpretability, and the lack of user-friendly statistical software. Overarching Goal*: This research is to develop innovative, flexible, generalizable and interpretable Bayesian methods to analyze large and complex real-world data.

Research Objectives and Approaches

Our proposal has two objectives. *Objective 1: Developing Scalable Hidden Markov Model with Informative Missingness*. Identifying patient phenotypes using EHR is of great importance to providing patients and carers with the chance of targeted intervention, better disease management, and efficient allocation of healthcare resources. We will develop a class model suitable for clinically monitoring patients' status via dynamic clustering, which involves two sub-models: (a) a multivariate continuous-time mixed effect hid- den Markov (multi-state) model describing the dynamic disease processes based on multiple longitudinal biomarkers, (b) a missing data model describing the state-dependent informative missingness. We will develop an efficient variational inference algorithm to approximate the posterior distributions for scalable inference. Simulation studies and applications to real data will be performed to evaluate the performance of the proposed methods compared to existing methods. A new R package will be developed to imple- ment the proposed model. This objective will be completed in Year 1. *Objective 2: Causal Inference with Latent Subgroups and Latent Confounders*. We will develop a Bayesian mixture model to estimate the causal effect on time-invariant clustering of multiple longitudinal responses and extend this method to causal mediation analysis with multiple longitudinal biomarkers and an end-of-study outcome. We will also develop a Bayesian

latent class approach to causal estimation with a time-varying treatment, time- varying latent biomarker clusters and an informative visiting process. Firstly, for the causal mixture model, we are interested in studying the changes in the group trajectory patterns of multiple outcomes by treat- ment. Secondly, we will extend the proposed approach to a mediation analysis where we are interested to estimate the natural direct and indirect effect of the latent group trajectory of multiple biomarkers on an end-of-study outcome. Secondly, we will extend the proposed approach to a mediation analysis where we are interested to estimate the natural direct and indirect effect of the latent group trajectory of multiple biomarkers on an end-of-study outcome. The causal estimates of interest are obtained via posterior pre- dictive inference by jointly specifying the latent mixture mediator models and the outcome model. As in Objective 1, we will develop an efficient inference algorithm to obtain the posterior distributions. Simu- lation studies and applications to real data will be performed to evaluate the performance of the proposed methods compared to existing methods. A new R package will be developed to implement the proposed model. This objective will be completed in Year 2. *Application*: This research is designed to address statistical challenges in analyzing longitudinal EHR data and answering pressing critical care research questions through collaboration. The PDF will lead the data analysis of the developed methods to study the effects of sustained interventions on key ventilation variables (e.g., tidal volume and driving pressure) on clinical outcomes for patients with acute respiratory failure (ARF) using the Toronto Intensive Care Observational Registry (iCORE). Our clinical collaborators, Drs. Fan and Munshi are leading researchers in critical care medicine specializing in ARF, who will facilitate access to iCORE data and provide clin- ical expertise on the application and implementation of the proposed research. *Relationship to NSERC Discovery Grant*: Dr. Lu is currently holding an NSERC Discovery Grant entitled *Modeling Longitudinal Data with Complex Structures*. The current proposal differs significantly from the projects described in the NSERC proposal in that the current proposal focuses on Bayesian methods related to dynamic clustering and causal inference using hidden Markov model with application to electronic health record data, which requires multi-disciplinary expertise as provided by Dr. Liu and Dr. Stephens in this application.

Training Environment

Mentorship Roles: The PDF will be supported to direct their own research, engage in the problem-solving and critical scientific appraisal of the work. All technical training uses a structured open-door policy, with team meetings between PDF and supervisors scheduled once a week and additional support offered by email or in person as needed. Monthly meetings between PDF, supervisors and mentor will support the ongoing research and professional training progress. Every 6 months, the PDF and Drs Lu, Liu and Stephens will formally discuss career goals and how training can be enhanced to meet specific goals. *Research Environment*: The proposed research will be carried out at Queen's University (Sep to Apr in Years 1 & 2) and the University of Toronto (May to Aug in Years 1 & 2). These are Canada's leading research-intensive universities with a global reputation for cutting-edge collaborative research intertwin- ing statistics, data sciences, medicine, and public health. The PDF will be provided with office space at the two host institutes and have access to seminars, training, scholarships, networking opportunities, col- laborations, and peer support through various initiatives including the Data Sciences Institute and Vector Institute at Toronto and the Canadian Cancer Trials Group at Queen's University.

Significance and Impact of the Proposed Research

The significance and impact are several folds: (a) this proposal facilitates new collaboration among fac- ulty members and the PDF from multiple disciplines and institutions with diverse expertise, (b) it will contribute substantially to advancing the area of Bayesian statistics by addressing several methodological gaps, including methods for dynamic clustering, causal inference and scalable computation, (c) it will fos- ter a mentorship environment with diverse expertise and experiences within and beyond the host institution that will be crucial to long-term success in the lead investigators' and the PDF's research career.

References

1. [1] LaurenJBeesleyandBhramarMukherjee.Statisticalinferenceforassociationstudiesusingelectronic health records: handling both selection bias and outcome misclassification. Biometrics, 78(1):214– 226, 2022.
2. [2] Rebecca A Hubbard, Jing Huang, Joanna Harton, Arman Oganisian, Grace Choi, Levon Utidjian, Ihuoma Eneli, L Charles Bailey, and Yong Chen. A bayesian latent class approach for ehr-based phenotyping. Statistics in Medicine, 38(1):74–87, 2019.
3. [3] Yang Ni, Peter Mu ̈ller, Maurice Diesendruck, Sinead Williamson, Yitan Zhu, and Yuan Ji. Scal- able bayesian nonparametric clustering and classification. Journal of Computational and Graphical Statistics, 29(1):53–65, 2020.
4. [4] YangNi,PeterMu ̈ller,andYuanJi.Bayesiandoublefeatureallocationforphenotypingwithelectronic health records. Journal of the American Statistical Association, 115(532):1620–1634, 2020.
5. [5] Yu Luo and David A Stephens. Bayesian inference for continuous-time hidden markov models with an unknown number of states. Statistics and computing, 31(5):1–15, 2021.