Canadian Statistical Sciences Institute
Graduate Student Enrichment Scholarships (GSES)
Cover Page
March 5, 2023

Submission Information

Date of submission	03/05/2023
GSES Program (AI, VS, NR)	NR
Proposed dates of support	September 2023 - September 2024
Requested amount of support	$15,000

Applicant information

Name of lead supervisor	Jessica Gronsbell
Email of lead supervisor	j.gronsbell@utoronto.ca
Home department and university of lead supervisor	Statistical Sciences, University of Toronto
Web page for lead supervisor	https://sites.google.com/view/jgronsbell/
Date of appointment of current position of lead supervisor	07/2020
Name of co-supervisor or sponsor	Linbo Wang
Home department and institution of co-supervisor or sponsor	Statistical Sciences & Computer and Mathematical Sciences, University of Toronto
Web page for co-supervisor or sponsor	https://sites.google.com/site/linbowangpku/

Student information

Name of student (write to ***To Be Recruited*** if unknown).	[redacted]
University and department	Statistical Sciences, University of Toronto
Degree program	PhD
Date enrolled or will be enrolled	September 2023 (offer to the program already accepted)

Jessica Gronsbell
Statistical Sciences
University of Toronto
j.gronsbell@utoronto.ca
https://sites.google.com/view/jgronsbell/

Professional Preparation

University of California at Berkeley, Berkeley, CA	BA, 2007
Harvard University, Cambridge, MA	PhD, 2012
Stanford University, Stanford, CA	Postdoc, 2017-2018

Professional Appointments

2022-present	**Mentor,** CANSSI Ontario STAGE program
2020-present	**Assistant Professor**, Department of Statistics, University of Toronto Cross appointed in Computer Science and Family & Community Medicine
2020-present	**Faculty Affiliate,** Vector Institute for Artificial Intelligence
2018-2020	**Data Scientist**, Alphabet's Verily Life Sciences, Cambridge, MA

Awards and Distinctions

2022	**Connaught New Researcher Award,** University of Toronto
2021-present	**AI Lead**, University of Toronto UTOPIAN Data Safe Haven
2019	**Best Paper in Knowledge Representation & Management,** IMIA
2016	**Gertrude Cox Award,** Honorable Mention**,** American Statistical Association
2016	**Young Researcher Award,** International Society of Nonparametric Statistics

Ten or Less Products Related to the Research Project

* - corresponding author, ** - trainee (**Note**: I left academia for industry from 2018-2020)

Informatics and Statistics

1. McCaw Z, Gao J**, Lin X, and **Gronsbell J***, *SynSurr: Leveraging a machine learning derived surrogate phenotype to improve power for genome-wide association studies of partially missing phenotypes in population biobanks*, Under review at **Nature Genetics.** **https://doi.org/10.1101/2022.12.12.520180.**
2. Yang S**, Varghese P, Stephenson E, Tu K, and **Gronsbell J*,** *Machine learning approaches for electronic health records phenotyping: A methodical review,* **Journal of the American Medical Informatics Association**, 30:2(2022), 367-381. (Editor's choice featured article.)

3. **Gronsbell J**, Liu M, Tian L, and Cai T, *Efficient Estimation and Evaluation of Prediction Rules in Semi-Supervised Settings under Stratified Sampling,* **Journal of the Royal Statistical Society, Series B**, 84:4(2022), pp.1353-1391I.
4. **Gronsbell J,** Hong C, Lie N, Lu Y, and Tian L, *Exact Inference for the Random-Effect Model for Meta-Analyses with Rare Events*, **Statistics in Medicine**, 39:3(2020), 252-264.
5. **Gronsbell J**, Minnier J, Yu S, Liao K, and Cai T, *Automated Feature Selection of Predictors in Electronic Medical Records Data,* **Biometrics**, 75:1(2019), 268-277.
6. Yu S, Ma Y, **Gronsbell J**, Liao K, Cai TT, Ananthakrishnan A, Gainer V, Churchill S, Szolovits P, Murphy S, Kohane I, and Cai T, *PheNorm: Full High-throughput Phenotyping with Denoised Normal Mixture Transformation*, **Journal of the American Medical Informatics Association**, 25:1(2018), 54-60.
7. **Gronsbell J** and Cai T, *Efficient Estimation of Prediction Performance Measures in Semi-Supervised Settings*, **Journal of the Royal Statistical Society, Series B**, 80:3(2018), 579-594.

Clinical Applications

8. Stephenson E, Yusuf A, **Gronsbell J,** Tu K, Mitiku T, Melamed O, Selby P, and O'Neill, B, *Disruptions in primary care among people with schizophrenia in Ontario, Canada during the COVID-19 pandemic*, **The Canadian Journal of Psychiatry**, p.07067437221140384 (2022).
9. Tu K, Kristiansson RS, **Gronsbell J**, de Lusignan S, Flottorp S, Goh L, Hallinan C, Hoang U, Kang S, Kim Y, Ling Z, Manski-Nankervis, J, Ng A, Pace WD, Wensaas K, Wong W, and Stephenson E, *Changes in primary care visits arising from the COVID-19 pandemic: An international comparative study by INTRePID, the International Consortium of Primary Care Big Data Researcher,* **BMJ open**, 12:5(2021), e059130.
10. Poole SF, **Gronsbell J,** Winter D, Nickels S, Levy R, Fu B, Burq M, Saeb S, Edwards MD, Behr MK, Kumaresan V, Macalalad AR, Shah S, Prevost M, Snoad N, Brenner MP, Myers LJ, Varghese P, Califf RM, Washington V, Lee VS, and Fromer M, *A holistic approach for suppression of COVID-19 spread in workplaces and universities*, **PloS one**, 16.8(2021), e0254798.

Ph.D. Students and Postdoctoral Fellows Supervised and Co-Supervised

Ph.D. Students	2 in progress

Type	years	Thesis or Project title	Current position
Ph.D.	2021-Present	*Statistical Methods for Integrating Genetic, Imaging, and Electronic Health Record Data*	N/A
Ph.D	2021-Present	*Statistical Methods for Prediction and Inference in settings of Complex Measurement Error and Misclassification*	N/A

LINBO WANG
Department of Statistical Sciences
University of Toronto
linbo.wang@utoronto.ca
https://www.statistics.utoronto.ca/people/directories/all-faculty/linbo-wang/

Professional Preparation

Peking University, Beijing, China	BS, 2011
University of Washington, Seattle, USA	Ph.D., 2016
Harvard University, Boston, USA	Postdoc, 2016-2018

Professional Appointments

2018-present **Assistant Professor**, Department of Computer & Mathematical Sciences, University of Toronto Scarborough

2018-present **Assistant Professor**, Department of Statistical Sciences, University of Toronto

2022-present **Assistant Professor**, Department of Computer Science, University of Toronto (cross-appointment)

2022-present **Mentor**, CANSSI Ontario STAGE program

2019-present **Faculty Affiliate**, Vector Institute

2019-present **Adjunct Assistant Professor**, Department of Statistics, University of Washington

Awards and Distinctions

2022-2023 **Guest Editor**, *Computation*, Special Issue on Causal Inference, Probability Theory, and Graphical Concepts

2022 **Outstanding Young Researcher Award**, International Chinese Statistical Association

2022 **Keynote talk**, Waterloo Conference in Statistics, Actuarial Science, and Finance

2022 **Doctoral Early Research Excellence Award**, University of Toronto Statistical Sciences
Awarded to my PhD student Sonia Markes for her thesis work

2021 **IMS Hannan Graduate Student Travel Award**, Institute of Mathematical Statistics
Awarded to my PhD student Ying Zhou for her thesis work

2021 **ICSA Student Paper Award**, International Chinese Statistical Association
Awarded to my PhD student Ying Zhou for her thesis work

2021 **Connaught International Scholarship**, University of Toronto
Awarded to my PhD student Xiaochuan Shi

2021 **Ontario Graduate Scholarship**, Ministry of Colleges and Universities, Ontario
Awarded to my PhD student Sonia Markes

2020 **Dean's Excellence Award for outstanding performance across the Faculty**, University of Toronto Scarborough

2020	**General Motors Women in Science and Mathematics Award**, University of Toronto *Awarded to my PhD student Ying Zhou for her thesis work*
2020	**Connaught International Scholarship**, University of Toronto *Awarded to my PhD student Dingke Tang*
2019	**Discovery Accelerator Supplements (DAS) Award**, Natural Sciences and Engineering Research Council of Canada
2018	**Nonparametric Statistics Section Student Paper Awards Finalist**, American Statistical Association *Awarded to my PhD student Xinyi Zhang for her thesis work*

Ten or Less Products Related to the Research Project

(Corresponding author* if not the first author; Student or postdoc collaborators)

1. **Wang L.**, Tchetgen Tchetgen E., Martinussen T., and Vansteelandt S. (2023+). Instrumental variable estimation of the causal hazard ratio (**with discussions**). *Biometrics*, to appear.
2. Lin, Z., Kong, D., and **Wang, L.*** (2023+). Causal Inference on Distribution Functions. *Journal of the Royal Statistical Society: Series B*, to appear.
3. Yu D., **Wang L.**, Kong D., and Zhu H. (2023+). Mapping the Genetic-Imaging-Clinical Pathway with Applications to Alzheimer's Disease. *Journal of the American Statistical Association*, to appear.
4. Tang D., Kong D., Pan W., and **Wang L.*** (2023+). Outcome model free causal inference with ultra-high dimensional covariates. *Biometrics*, to appear.
5. Zhou Y. , Tang D., Kong D., and **Wang L.*** (2023+). The Promises of Parallel Outcomes. *Biometrika*, minor revision.
6. **Wang L.**, Meng X. , Richardson T. S., and Robins J. M. (2023+). Coherent causal modeling of longitudinal effects on dichotomous outcomes. *Biometrics*, to appear.
7. Kong D., Yang S., and **Wang L.*** (2022). Identifiability of causal effects with multiple causes and a binary outcome. *Biometrika*, 109(1): 265-272.
8. Yin, J., Markes, S., Richardson, T. S., and **Wang, L.*** (2022). Multiplicative effect modelling: the general case. *Biometrika*, 109(2), 559-566.
9. Bilodeau, B., **Wang, L.**, and Roy, D. M. (2022). Adaptively Exploiting d-Separators with Causal Bandits. (**oral presentation**). Advances of Neural Information Processing Systems (NeurIPS 2022).
10. **Wang L.**, Zhang Y., Richardson T.S. and Robins J. M. (2021). Estimation of local treatment effects under the binary instrumental variable model. *Biometrika*, 108(4), 881-894.

Ph.D. Students and Postdoctoral Fellows Supervised and Co-Supervised

Ph.D. Students	1 graduated, 8 in progress
Postdoctoral Fellows	1 completed

Type	years	Thesis or Project title	Current position
Ph.D.	2018-2020	Binary Regression Models: From Association to Causation	Senior Applied Scientist, Microsoft

Ph.D.	2018-2023	The Promises of Parallel Outcomes	Ph.D. student, University of Toronto, Statistical Sciences (Next: Postdoctoral fellow, University of Pennsylvania & Assistant Professor (tenure track), University of Connecticut)
Ph.D.	2018-2023	Fighting noise with noise: Causal Inference with Many Candidate Instruments	Ph.D. student, University of Toronto, Statistical Sciences (Next: Postdoctoral fellow, Johns Hopkins University)
PDF	2019-2022	Causal Inference with Massive and Complex data	Assistant Professor (tenure track), College of Business, University of Texas at San Antonio
Ph.D.	2020-present	Multiplicative effect modelling	Ph.D. student, University of Toronto, Statistical Sciences
Ph.D.	2021-present	Causal effect estimation for multiple treatments	Ph.D. student, University of Toronto, Statistical Sciences
Ph.D. (visiting)	2023-present	Gene expression prediction based on neighbour connection neural network utilizing gene interaction graphs	Ph.D. student, University of Chinese Academy of Sciences
Ph.D. (visiting)	2023-present	Study on Doubly Robust Estimation in Causal Inference	Ph.D. student, Beijing Normal University
Ph.D.	Starting 2023	Z-estimation under correlated outcome and covariate measurement error	Master student, University of Toronto, Statistical Sciences

Intended Research Program + Planned Interactions with Supervisors

Research Program

Electronic health records (EHRs) are a rich source of data for generating real-world evidence of treatment effectiveness in settings where clinical trials are underpowered, infeasible, or unethical. However, inferring a treatment effect from EHR data is methodologically challenging due to the presence of both (i) measurement error (or misclassification) in the treatment and response and (ii) unmeasured confounding.

With respect to (i), treatment and response information are often not readily available in EHRs. For example, responses such as disease remission and progression must be inferred from information across clinical notes and various structured data elements (e.g., labs, diagnosis codes). Similarly, treatment information is not fully captured by prescription codes. Consequently, treatment and response variables are derived from machine learning (ML) methods that combine multiple EHR data elements. The ML-derived variables are at best strong surrogates for a patient's true treatment and response as they are inevitably prone to measurement error. Moreover, with respect to (ii), confounders such as economic status, physical activity, and genetic markers, are often not recorded due to the observational nature of EHR data.

While numerous statistical methods for treatment effect estimation have been proposed to address either measurement error or unmeasured confounding, there is a paucity of methods that simultaneously address both issues. To fill this gap, we propose to bridge recent developments in semi-supervised learning and causal inference to enable valid treatment effect estimation in EHR-based studies. Specifically, we will develop methods for treatment effect estimation that leverage a massive EHR dataset subject to measurement error and unmeasured confounding together with a much smaller *gold-standard and deconfounded dataset* that contains the true treatment, response, and confounders.

Our project seeks to answer the following questions:

1. In what settings and how much can we improve over classical approaches that rely only on gold-standard and deconfounded data in terms of statistical efficiency?
2. How well do our methods perform on real-world EHR data?

For (1), we will establish theoretical results supporting our proposed method (e.g.,, characterization of semi-parametric efficiency). For (2), we will utilize the MIMIC-III benchmark dataset for causal inference. MIMIC-III contains EHRs on intensive care unit admissions for 38,645 adults at the Beth Israel Deaconess Medical Center (BIDMC) in Boston.

The causal inference benchmark contains both EHR data and gold-standard and deconfounded datasets from 3 clinical trials conducted at BIDMC.

Planned Interactions

We will provide the PhD trainee ([redacted]) with exceptional statistical training in both semi-supervised learning (Dr. Gronsbell) and causal inference (Dr. Wang), which are our complementary areas of expertise. We will also guide [redacted] in the analysis of EHR data (Dr. Gronsbell).

Throughout her training, [redacted] will meet with us weekly to discuss her research progress. She will also attend Drs. Gronsbell and Wang's monthly joint group meeting to gain further exposure to statistical research. Our planned interactions are detailed in the year-by-year training plan below.

- **Year 1**: Focus on completing coursework for the comprehensive examination. [redacted] will also begin a literature review of causal inference with EHR data. Her findings will be synthesized into a tutorial paper for publication at a clinical/informatics venue (e.g., *JAMIA*). This paper will provide [redacted] with necessary background on EHRs and enhance her ability to explain statistics to applied researchers.
- **Year 2**: Complete the tutorial paper and begin the first dissertation paper. [redacted] will critique an existing causal inference method (e.g., Gan et al 2021) for the oral comprehensive exam and then develop our proposed extension. [redacted] will also assist a clinical collaborator of Dr. Gronsbell with a simple EHR project to gain experience with real data.
- **Year 3**: Present research progress at a statistical conference (e.g., Statistical Society of Canada). [redacted] will also submit her paper from Year 2 to a top statistical journal such as *JRSS:B* and release a corresponding R package.
- **Year 4**: Implement an extensive application of her method from Year 3 with MIMIC-III data for publication in *JAMIA*. [redacted] will continue to present research at statistical and informatics conferences and begin another statistical methods paper.
- **Year 5**: Submit the papers from Year 4 and complete an additional methods/applied statistical paper. [redacted] will also receive support in obtaining an academic or industry position.

Support Letter

Linbo Wang
Assistant Professor
Department of Statistical Sciences, University of Toronto
Department of Computer and Mathematical Sciences, University of Toronto Scarborough

Dear Dr. Jessica Gronsbell,

I am writing this letter to express my full support for your proposed co-supervision of [redacted], an incoming Ph.D. student in the doctoral program in the Department of Statistical Sciences at the University of Toronto. I have had the pleasure of working with you closely over the years through research collaboration, co-supervision, and joint group meetings, and have been consistently impressed by your expertise and commitment to teaching and research.

Given your extensive experience in statistical methodology and applications, I am confident that you possess the necessary qualifications to effectively guide Ph.D. students and help them excel in their careers. Your research in semi-supervised learning and applications to electronic health records has produced significant contributions to the field and has provided a solid foundation for mentoring future scholars in the field. I am confident that our complementary areas of expertise and research interests would provide a holistic understanding of the challenges facing the field, which would enrich the student's research.

I am pleased to hear that you have identified [redacted] as a potential Ph.D. student to co-supervise. I have reviewed [redacted] academic record and found her to be an outstanding candidate for the program. [redacted] has a great educational background, having graduated with an impressive GPA from the University of Waterloo, one of the top institutions in Canada for statistical science. She is also currently a master's student in our department and is doing extremely well. Additionally, I have had the pleasure of interacting with [redacted] on a few occasions and have found her to be highly organized, detail-oriented, and enthusiastic about her work. I am excited that she has accepted the offer to join our Ph.D. program and am looking forward to working with her more closely in the next few years.

Therefore, I wholeheartedly endorse your co-supervision of [redacted] for our Ph.D. program in Statistical Sciences. I am confident that with our mentorship, [redacted] will thrive academically and make valuable contributions to the field.

Best regards,
Linbo Wang

EDI

We agree to satisfy the following CANSSI EDI requirements:

- [redacted] will take part in CANSSI EDI which is CANSSI's annual program of activities related to equity, diversity and inclusion and involves completion of at least one CANSSI EDI event.
- Both Dr. Gronsbell and Dr. Wang will also take part in at least one activity in CANSSI EDI.

Additionally, we are committed to maintaining an inclusive multicultural and gender-diverse environment. While statistics is a field that traditionally suffers from a lack of diversity in race and gender, Drs. Gronsbell and Wang are recognized as leading diverse research groups at U of T. Moving forward, we will adhere to the following standards to maintain an inclusive and equitable environment in our research groups:

- Annual assessment of all EDI policies established by the University of Toronto (https://hrandequity.utoronto.ca/inclusion/edi-at-u-of-t, https://hrandequity.utoronto.ca/policy/).
- Group participation in an annual EDI program offered by the University of Toronto on subjects including anti-racism, anti-oppression, and EDI vocabulary and frameworks.
- Training on indigenous cultures and unconscious bias in groups.
- Distributing networking and professional development opportunities to trainees from diverse backgrounds in an equitable manner.

Budget

If successful, the funds from CANSSI will be used to cover most of the departmental contribution for a PhD student required by our department ($20,000/year). The remaining $5,000 tuition will be covered by the NSERC Discovery Grants currently held by Dr. Gronsbell ($18,000/year) and Dr. Wang ($30,000/year). The additional costs and source of funds for this project beyond the student funding package are described below.

- All computing costs are covered by our affiliations with the Vector Institute for Artificial Intelligence which provide our students with high-performance computing resources.
- The MIMIC-III is an openly available dataset which is free to use.
- Publication costs and travel to conferences will be covered by Dr. Gronsbell and Wang's start-up funds. We estimate about $3,500 in conference travel in Years 2-5 and open access publication costs of approximately $3,000 per publication in informatics journals.

As inflation gets more severe, we also want to note that there will be opportunities for [redacted] to obtain research assistantships supporting clinical researchers to top-up her salary if she desires. The detailed budget for her entire PhD study is described below.

Year	**Cost**	**Source of Funds**
1	Tuition	$15k - CANSSI $5k - Gronsbell + Wang
2	Tuition Conference Travel Open Access Publication	$20k - Gronsbell + Wang $3.5k - Gronsbell + Wang $3k - Gronsbell + Wang
3	Tuition Conference Travel	$20k - Gronsbell + Wang $3.5k - Gronsbell + Wang
4	Tuition Conference Travel	$20k - Gronsbell + Wang $3.5k - Gronsbell + Wang
5	Tuition Conference Travel Open Access Publication	$20 - Gronsbell + Wang $3.5k - Gronsbell + Wang $3k - Gronsbell + Wang