Form I: BTI Internship Program Mentor Application

Mentor		
Name Lillian Sung	Institution The Hospital for Sick Children	Department Pediatrics

A.2: Co-mentor		
Name George Tomlinson	Institution University Health Network	Department Medicine

A.3: Research proposal

Title
Identifying bothersome symptoms in pediatric cancer patients

Background: This program is focused on using big data to improve supportive care for children with cancer. Over the last few decades, impressive gains in survival have been made and now, more than 82% of pediatric cancer patients will be long term survivors.[1] These survival gains have been, in part, attributable to the provision of intensive therapies. However, these therapies frequently result in toxicities including severely bothersome symptoms such as pain, fatigue, nausea and changes in hunger.[2] These side effects of therapy are important as they result in morbidity, prolong hospitalization, compromise chemotherapy delivery, reduce quality of life and increase health care utilization.

There is an increasing number of large, complex datasets in pediatric cancer research that could provide insight into predictive patterns of toxicities including symptoms. This insight could lead to different therapeutic approaches to manage patients, such as identification of high-risk patients meriting enhanced monitoring or prophylactic strategies. However, in order to use data in predictive models, an "electronic phenotype" is required.

We previously developed the Symptom Screening in Pediatrics Tool (SSPedi).[3, 4] SSPedi is a self-reported, 15-item symptom screening tool for children receiving cancer treatments. It is reliable and valid in English-speaking children 8-18 years of age receiving cancer therapies.[2] Mini-SSPedi has face and content validity and is available for pediatric self-report for children 4-7 years of age. Across multiple studies, SSPedi and mini-SSPedi have been utilized by over 1000 children at The Hospital for Sick Children (SickKids), thus providing a gold standard. This proposal will use this gold standard to develop an electronic phenotype for the 15 symptoms included in SSPedi and mini-SSPedi. In other words, we will use data in the EHR to identify if patients experience the 15 symptoms without the need to rely on patient self-report. Once developed, it may be used for future studies focused on determining effective approaches to prevent and manage symptoms. We also have manually abstracted whether symptoms are documented or treated in the electronic health record (EHR) as part of other projects, providing additional important information.

Aims: To develop electronic phenotypes for 15 symptoms in pediatric cancer patients using data in the SickKids EHR.

Research Methodology: The data sources will be the SickKids EHR (Epic) and the divisional haematology/oncology database that captures all pediatric patients with cancer. We will include pediatric patients with cancer or hematopoietic stem cell transplantation recipients who have received any cancer treatments after July 1, 2018 and later (date Epic was instituted) and who were at least 4 years of age when trying to predict if they have one of the 15 symptoms.

To develop the algorithm, the pediatric cancer EHR cohort will be linked to our local SSPedi/mini-SSPedi data using hospital medical record numbers (MRN). For each patient enrolled on a SSPedi or mini-SSPedi study, we have the patient reported symptom ratings. In addition, for those with severely bothersome symptoms, we have abstracted whether there was any documentation of that symptom or any intervention provided ± 2 days of reporting.

Using the linked data, for each symptom, we will develop models that predict four targets (outcomes),

namely severely bothersome symptoms, any bothersome symptoms, documentation of symptoms or intervention provision. The features (potential predictors) will include vital signs, laboratory and radiology tests, medications, blood product utilization, procedures, admission/discharge/transfer data and notes (natural language processing). We will use regularized logistic regression, random forest and neural nets and will vary the length of the feature window. Metrics evaluated will be area under the receiver operator curve, precision-recall curve and a measure of loss specific to the model. The model will be trained on 60% of the patients chosen randomly with hyperparameters selected based on a validation set (20%). Metrics will be described in the held-out test set (20%). Once developed, we will apply the model to patients outside the SSPedi/mini-SSPedi cohort to describe symptoms in patients who did not participate in a SSPedi study at multiple time points relative to cancer diagnosis date.

Training Objectives: Training objectives will be to gain competencies in managing large clinical datasets (EHR), developing predictive models and creating and interpreting machine learning algorithms.

Environment including Mentorship Plan

Dr. Sung will be the Principal Mentor. She is a Pediatric Oncologist with a PhD in Clinical Epidemiology. Dr. Sung has a robust research program and has received continuous grant funding from the Canadian Institutes of Health Research (CIHR), Canadian Cancer Society Research Institute and the National Institutes of Health (R01). She has spent since September 15, 2019 working with the Biomedical Informatics Research group at Stanford to develop a machine learning research program focused on supportive care. She has hired a machine learning specialist as part of her research staff who will also provide support.

Dr. Tomlinson has a PhD and is a Scientist in the Department of Medicine, UHN. He will be the biostatistician co-mentor. He is the director of the Biostatistics Research Unit at UHN (https://thebru.ca). He has 25 years of collaborative experience across many fields of medicine and epidemiology. He is the primary statistician on numerous completed and ongoing clinical trials. He has direct experience in mentoring more than 60 biostatisticians or clinical epidemiologists learning biostatistical skills.

Mentorship Team: Drs. Sung and Tomlinson have been close collaborators across many research projects for almost 20 years. They co-coordinate HAD5313H - Advanced Design and Analysis Issues in Clinical Trials, an advanced graduate course for MSc and PhD students in Institute of Health Policy, Management and Evaluation (IHPME) at the University of Toronto.

Both Drs. Sung and Tomlinson are primary supervisors for PhD students who have thesis projects focused on machine learning, thus, enriching the environment for the intern. There are resources at SickKids and UHN that will also facilitate training. The SickKids Research Training Centre and Faculty Office offers educational programs centered on skill development and career paths. Training is also available through the Centre for Computational Medicine (CCM), which is a core facility within the SickKids Research Institute. The center offers weekly courses in the fall and spring on various bioinformatics tools. Research-related rounds are available at SickKids, including weekly Child Health Evaluative Sciences seminars, Haematology/Oncology Divisional Rounds, Paediatric Grand Rounds and journal clubs. Computing support will be offered by SickKids Research Information Technology, which provides core computing resources, services and infrastructure operated in accordance with institutional security and privacy polies, and federal and provincial requirements. The intern will have access to the High Performance Computing Facility. At UHN, there are weekly rounds on health economics or clinical epidemiology and the intern will be invited to attend monthly biostatistics rounds organized by the group of statisticians that Dr. Tomlinson manages at the Biostatistics Research Unit. In order to enhance collaboration and peer mentoring, the intern will spend time each week (potentially virtually, pending COVID restrictions) at UHN interacting with the other biostatisticians (minimum one day per week). This would provide an opportunity to be involved in biostatistics activities including consultations and ad hoc discussions on methods.

Necessary Intern Background/Strengths: Major requirement is enthusiasm to learn and to be an active participant in developing novel machine learning approaches to pediatric cancer EHR data. Knowledge of R,

Python and SQL are an asset.

Alignment with OICR's Strategic Plan: This research aligns with the following goals of OICR: to perform cutting-edge translational cancer research, mobilize Ontario research strengths around key cancer priorities and global leadership in precision oncology. Symptom control has been an Ontario priority for over a decade, with widespread adoption of routine collection of the Edmonton Symptom Assessment Scale in clinics across the province (the adult equivalent of SSPedi and mini-SSPedi). Electronic phenotyping is a cutting-edge priority in machine learning and this proposal will allow the mobilization of Ontario research strengths in machine learning and artificial intelligence.

References:

1. Canadian Cancer Society's Steering Committee on Cancer Statistics. *Canadian Cancer Statistics.* Toronto, ON: Canadian Cancer Society;2011.
2. Dupuis LL, Johnston DL, Baggott C, et al. Validation of the Symptom Screening in Pediatrics Tool in Children Receiving Cancer Treatments. *Journal of the National Cancer Institute.* 2017.
3. O'Sullivan C, Dupuis LL, Gibson P, et al. Refinement of the symptom screening in pediatrics tool (SSPedi). *Br J Cancer.* 2014;111(7):1262-1268.
4. O'Sullivan C, Dupuis LL, Gibson P, et al. Evaluation of the electronic self-report Symptom Screening in Pediatrics Tool (SSPedi). *BMJ supportive & palliative care.* 2016.