OVERVIEW

The goal of segmentation is to stratify the population into mutually exclusive subsets according to health status, health care needs and priorities for the purpose of health system management, which aims to better resource, organize and proactively deliver care for similar groups of people with similar needs. This approach allows consideration of how to best respond to different population segments in different ways. Various approaches have been developed to segment the population including both expert-driven and data-driven approaches to clustering for specific conditions or diagnostic codes, and to a lesser extent, risk prediction models. Despite the enthusiasm around the potential for population health management and the use of data to guide these approaches, there remains some significant gaps that hinder its applications. Firstly, very few of these models are validated or tested in populations in advance of being implemented. For example, although approaches have been applied to Canadian data, they have not been done with the intention of integrating for population health management and testing with decision-makers. This has significant implications on the way health services and resources are managed and could affect downstream costs and health outcomes. The second major gap is that most of these models focus on utilization for inputs (e.g. hospitalization records), which can significantly impact their ability to fully characterize risk in the population, particularly through characteristics that are antecedent to illness, such as socioeconomic and health behavioural determinants that are not available in traditional health care utilization data. This limits the ability to inform population health approaches, which aim to lessen disparities in health delivery and improve health over an entire population,[1] not only for those seeking certain types of care. The final gap is that very few people know what to do with the outputs of these segmentations (e.g. population clusters) in an explicit way that informs population health management. By rigorously testing existing and new analytic approaches to segmentation and studying the impact in a decision-making organization, there is an opportunity to better use population level data to inform health system management.

Goal: The goal of this grant is to compare different (existing and new) analytic approaches to executing population segmentation in the Canadian context.

The **objectives** are to:

1. Analyze and compare a provincial and rostered population according to existing segmentation approaches that are expert-driven, data-driven and predictive;
2. To validate the different segmentation approaches for health outcomes (i.e. health care costs, high cost users, multiple chronic diseases, premature mortality, and avoidable hospitalizations);
3. Present and test the different approaches to a decision-making team, responsible for population health management.

Working closely with a knowledge user partner, Trillium Health Partners (THP), we will demonstrate the applied impact of this work through a robust partnership and knowledge translation strategy.

BACKGROUND AND RATIONALE

Managing Populations

There is increasing pressure to implement population health management for health care. Population health management is a widely used term with various definitions; however, for our purposes, it is an approach that describes implementing health care strategies that are tailored to the health care needs of the population.[2] We conceptualize needs based on the Risk Adjustment Framework to Primary Care[3] which posits that health care needs are determined by not only morbidity types and mix, but also demographics, health behaviours, patient preferences, and socioeconomic factors that either enable or detract from people receiving health care services. The term population is often misused and

misinterpreted in the context of population management. Specifically, the term should encompass whole populations (i.e. defined by geography, insurance coverage or attribution), and not only those with specific illness or needs.[4] The segmentation approach will serve to identify sub-populations within the population that are meaningful and actionable from a health care perspective.

Population segmentation approaches

An analytic approach to carry out population segmentation involves using population-level data to stratify the population into groups with similar health care needs. Strategies are then expected to align with these segments. There have been three broad approaches to carry out population segmentation: expert/consensus-driven approaches, data-driven approaches, and predictive approaches. We will describe these three approaches and provide a rational basis of each type being applied in this grant:

Expert/consensus-driven: These approaches involve bringing together clinical and health care experts and using consensus approaches to define the list of important diagnoses or utilization patterns and then using analytic approaches that rank or weight these sub-groups. There is some variation in the approaches, for example, some do this hierarchically, whereas others weight based on a cost criterion, and some approaches combine both. This is then used to guide the analysis of data on the population to form meaningful clinical clusters. The Johns Hopkins Adjusted Clinical Groups (ACGs)® system is the most popular expert/consensus-driven approach and was generated using existing data from medical claims, electronic medical records, and demographics. In 2017, the Canadian Institutes of Health Research (CIHI) developed a similar Population Grouping Methodology using person-level hospitalization, ambulatory care, continuing reporting system, and physician billing data and built from 226 expert-informed clinical groupings.[5]

Data-driven: Data driven-approaches typically involve clustering or other post-hoc approaches that use patterns in utilization data to generate clusters that will then go on to represent population segments. In some cases, these approaches are combined with the expert driven approaches in a second step. There are no routine data-driven approaches in wide-spread use in Canada. As such, we have recruited an expert in these approaches to our study team to implement two novel clustering-based approaches. First, we will apply a density-based clustering approach, recently shown to produce the most clinically distinct and greatest variation in health care costs when applied to Medicare data compared to other clustering approaches.[6] Secondly, we will apply an approach led by our US collaborator Dr. Udell, known as generalized low rank models,[7] which have started to be applied to identify meaningful clusters in varied health data.[8] This clustering technique builds from the more commonly known Principal Component Analysis (PCA) but extends to the type of data we encounter in transactional population health data. These approaches have proven to be a powerful framework for clustering or embedding examples in a low dimensional space, based on partial observations of heterogeneous data, including real, Boolean, interval, ordinal, or categorical observations. Dr. Udell has developed a user-friendly open source implementation[9] and has further developed methodology for sparse data.[10]

Predictive: Predictive and forecasting approaches are essential for effective health planning and decision-making.[11] These approaches allow us to measure and forecast the impact of a proposed policy or program on future disease burden and related costs. The challenge is that existing approaches often fail to account for the influence of upstream risk factors, including health status, behavioural and social factors. Additionally, decision-makers are challenged with how best to use existing data to estimate the impact of prevention approaches on health care utilization or potential cost savings. By applying prediction tools that accurately capture all attributes of a person, not just their previous health care utilization or diagnoses, we can better understand how the broader health system can contribute

towards overall health system performance and sustainability.[12] Developing these novel analytic approaches will help decision-makers weigh the potential costs and benefits of adopting various preventive strategies and their impact on downstream costs. Prediction models can be developed and validated to predict hospital or health care utilization and costs. These statistical or machine learning models are then applied to a population to predict the likelihood of utilization; the probabilities or scores generated from the prediction models are then used to discriminate the population by groups that are likely to be higher users or have specific utilization compared to lower probabilities where this is unlikely. The NPA has led the development and validation of the High Risk Population Risk Tool (HRUPoRT),[13] which is one example of a predictive model that can be used for population stratification. Specifically, the predictive probabilities can be used to identify target groups and interventions to optimize population health outcomes (reduced incidence or lower costs), which we have demonstrated in our previous work.[14 15] We will also apply the Combined Predictive Model (CPM)[16] which is a predictive approach to segmentation developed in the UK.[17] This approach is based on combining predictive models for hospitalization and readmission on a broader range of data. Finally, both the Johns Hopkins Adjusted Clinical Groups (ACGs)® and the CIHI Population Grouping Methodology have predictive models developed, but they are not typically used. We will also explore the use of these models for segmentation processes.

A focus on utilization and billing data: missing the social and behavioural determinants of risk

Overwhelmingly, the inputs to the existing segmentation approaches are based primarily on health claims or billing data. This represents a major gap in segmentation due to the fact that it only segments the population on health care utilization patterns and illness, and does not explicitly consider population health factors such as demographics, socioeconomic, and environmental factors.[18 19] This limits the ability to consider factors that fully capture the spectrum health to illness,[20] and will limit the ability to discriminate high or lower risk groups within a homogenous clinical population (for example low or high socioeconomic (SES) populations within a diabetes population). The number of attributes that a data set has is referred to as the dimensionality. Operations and machine learning researchers often refer to this type of data as low dimensional data.[21] Health care data with multiple transactions and ICD codes contains a lot of information and data which are related to each other, which increases computational requirements and interoperability. For this reason, clustering approaches and hierarchical implications have been a primary strategy to gain information from this data. One challenge with this approach is that by applying hierarchical and simple data reduction techniques, information may be lost. Embracing the dimensionality requires newer techniques.

RESEARCH APPROACH/METHODS

Data sources

We intend to use a range of data sources to apply our segmentation approaches as described below: We will apply various segmentation approaches on a range of population-level data sources in Ontario and within an attributed population. ***Ontario Health Insurance Plan (OHIP):*** OHIP covers all Ontario residents as a single-payer health insurance system. We will use encrypted OHIP numbers (unique identifiers) to link data between health administrative databases and OHIP physician claims file for ambulatory physician visits and imaging. ***Canadian Institute for Health Information (CIHI):*** CIHI Discharge Abstract Database and Same Day Surgeries collect demographic, administrative, and clinical data on hospital discharges and same day surgeries, which are received from acute care facilities, health/regional authority, or ministry of health (depending on the province). We will use this to obtain data on hospitalizations and imaging. ***National Ambulatory Care Reporting System (NACRS):*** This captures data on all hospital-based and community-based ambulatory care, collected from specific facilities, regional health authorities, and ministries of health. This will be used to ascertain data on

emergency department (ED) visits and imaging. ***Registered Persons Database (RPDB):*** The RPDB captures demographic information on all individuals who received a health card number in Ontario, and updates home addresses when an OHIP card is issued or renewed, or with address changes registered in EDs or hospitals. This will be used to identify demographic information and censoring when subjects move out of Ontario. ***Health care costs:*** From 2003 onwards, comprehensive health care costs are available for all major sectors of health care spending: inpatient hospitalizations, physician visits, complex continuing care, long-term care, home services, assistive devices and pharmaceuticals.[22] We have successfully applied these methods to estimate attributable costs for other conditions.[22-25] ***Population Health Surveys:*** Canadian Community Health Survey (CCHS) cycles from 2015/16 will be accessed through Statistics Canada Research Data Centers. These data allow for in-depth analysis on variables not available through administrative data such as individual-level income, education, food security, health behaviours, life stress, perceived well-being and other health factors.
CCHS is administered by Statistics Canada and gathers up-to-date, cross-sectional data on the distribution of health determinants, outcomes, and health care use across Canada.[26] CCHS data are representative of 98% of the Canadian population aged ≥12 years living in private dwellings at national and provincial levels (response rates>75%).[26] Survey methodology is detailed elsewhere[27]. CCHS data linked with health administrative data will be used to conduct sub/sensitivity analyses. ***Social Assistance Data - Ministry of Community and Social Services:*** We will use administrative caseload data provided by the Ontario Ministry of Community and Social Services, linked at ICES, to identify all users of Ontario's two major social assistance programs (Ontario Disability Support Program, Ontario Works).[28 29] This will be used for data on disability and social assistance use (e.g., benefit unit pay details, income and deductions, skills and training) as key covariate information. **Office of the Registrar General-Deaths File (ORG-D):** This database contains all deaths registered in Ontario since 1990 and is linked to ICES data holdings.[30] ***Corporate Provider Database (CPDB):*** This database is maintained by the Ministry of Health, and contains information on physician's age, sex, year of graduation, physician specialties, and postal code of practice. This will be used for data on provider-level factors as key covariate information. ***Population Demographic Files:*** We will use the Citizen and Immigration Canada (CIC) data to identify immigrants and refugees in the population; denominator data will be taken from Statistics Canada census files. ***SES:*** We will attribute census-derived area-level SES data to individuals, using the 2016 census nearest to year of population segmentation and the smallest available geographic census area (i.e., dissemination area (DA)). Additionally, for a subpopulation of the cohort, we will further describe SES trends according to the Ontario Marginalization (ON-Marg) Index, which measures area-level dependency, material deprivation, residential instability, and ethnic concentration. ***Geographic identifiers:*** All data are geo-coded at the postal code level and will be summarized using postal code conversion files according to relevant geographies within Ontario.

Feasibility: All of these databases are accessible to the study team and are governed by strict policy procedures and access regulations. We will aim to use 2018-2019 data, which will be complete for the full population. We will use the most recent version of the data as it becomes available. All Statistics Canada Data will be accessed through the Research Data Centre at UofT.

Sex and Gender considerations

The population segmentation approaches are likely to have important sex and gender differences. From a biological perspective, there are sex-related differences in the way that risk factors operate. For example, the relationship between chronic diseases and upstream population health determinants operate differently between men and women. There are also sex-related differences in hormonal and genetic function that contribute to the risk of sex-specific diseases. From a sociological perspective,

behavioural risk factors related to chronic diseases (e.g. food, physical activity) as well as health resource utilization patterns (screening, preventive care, frequency/type of physician visits) differ according to sex. Additionally, the health differences by gender are related to differences in the distribution of behavioral and social risk factors, including tobacco use, alcohol consumption, and social engagement. Further, there is evidence that women use health care services differently than men (greater use and more preventive service uptake). We have access to sex information through linkage with demographic information from health cards. Acknowledging these important differences, we will ensure that sex differences are considered throughout the project. We unfortunately do not have information on gender and will not be able to measure gender differences directly. However, we will include any caveats and important gender considerations through the interpretation of our results.

Objective-specific methods:

Objective 1 *Analyze and compare a provincial and rostered population according to existing segmentation approaches.*

Expert-driven approaches: We will apply the Johns Hopkins Adjusted Clinical Groups (ACGs)® approach system,[31] which has been routinely applied to the data sources proposed in this grant for the past 10 years, largely led by our Co-Investigator Dr. Peter Austin.[32-34] Co-I, Dr. Robert Reid was a pioneer in using ACG for case-mix in Canada[35-37] and additionally has extensive experience with the use of this system in the Kaiser health system. Briefly, we will use both the Adjusted Diagnostic Groups (ADG), which result in 32 conditions based on the expert opinions regarding condition duration, severity, diagnostic certainty, etiology and specialist care. We will apply both the clusters and the validated mortality Risk Score based on the ADGs, validated by Co-I Dr. Austin,[34] to segment the population. Recently CIHI created the Population Grouping Methodology[5] which similarly uses health administrative data from hospitalizations, day surgery, emergency department visits, physician claims, Long Term Care (LTC) and Complex Continuing Care (CCC) to output condition groupings known as the Health Profile Group (HPG). The model was developed in Canada; however, due to its newness has not been as widely used. This grant offers an opportunity to formally test across important outcomes and a range of other approaches.

Data driven approaches: Data driven approaches form clusters by defining the distances between subjects based on the combined values of their measured characteristics. We will apply a density-based approach to all utilization data making use of the ordering points to identify the clustering structure (OPTICS) algorithm,[38 39] which was recently tested for segmentation in a US population by Yan et al.[6] This approach was shown to be more robust and flexible with respect to cluster length and size and did not require prior specification of the clusters and therefore truly represents a data driven approach.

Prediction Approaches: We will apply a predictive model for population segmentation that incorporates behavioural and socioeconomic risk factors in the modelling. This will capture aspects of health care needs that are often antecedent to illness, thus offering a unique advantage to other approaches to population segmentation that focus solely on morbidity and health care use. We will apply the validated High Risk Population Risk Tool (HRUPoRT),[13 40] which is one example of a predictive model that can be used for population stratification that includes population data inputs that are lacking from all other approaches, such as sociodemographic, behaviours and more nuanced measures of SES factors (i.e. food security). HRUPoRT will be applied to routinely collected risk factor information available in the CCHS accessed at Statistics Canada Research Data Centres. Importantly, tools built on these widely accessible survey data allow decision-makers – who can access

these data (including regional data) – to perform segmentation for their population of interest. All models will incorporate Statistics Canada survey weights to accurately reflect population demographics and account for complex sampling design and non-response bias. Variance estimates will be calculated with bootstrap weights using Balanced Repeated Replication.[41 42]

Data preparation will be done using SAS. R (rtnse and cluster packages), Julia (programming language) will be used for clustering approaches, and a combination of SAS and R will be used for prediction.

Objective 2. *Validate the different segmentation approaches for health outcomes.*

Our goal for objective 2 is to understand how the different segmentation approaches segment the population on population health outcomes as well as health care utilization. We will then validate the different segmentation approaches for health outcomes: overall mortality, premature mortality, total health system expenditures, high cost users, avoidable hospitalizations, and multiple chronic diseases.

Health outcomes

We have chosen to evaluate our segmentation approaches across six complementary yet related health system outcomes. They were chosen to reflect broad measures of health system functioning and population health and thus allow us to provide an understanding of how the different segmentation approaches would inform how to resource, organize and proactively deliver care for groups of people with similar needs. Segmentation approaches will be calculated for fiscal year 2018/19, starting on April 1 2018 as the index date. Health outcomes will be calculated for 1-year following the index date.

Overall and premature death: We will calculate both overall (death at any age) and premature mortality rates (deaths before the age of 75) within 365 days of the index date.
Total health system expenditure: Annual health care spending will be calculated using administrative data from all health care sectors using established person-based costing methods developed by Co-I Dr. Wodchis and colleagues at ICES. Specifically, comprehensive person-centered health care cost estimates [43] and an assembly of disease cohorts [44] will be created. Health care costs are available for all major health care sectors: inpatient hospitalizations, physician visits, assistive devices, and pharmaceuticals (for certain patients, e.g. seniors), according to well-established methodology.[45]
High costs users: This outcome is an alternate conceptualization of costs by focusing on the highest users of health care services (top 5%), a meaningful endpoint for health system decision-makers.[19 43]
Avoidable hospitalizations: We will estimate avoidable hospitalizations as a complementary health system outcome. Ambulatory care sensitive conditions (ACSCs) are a set of conditions for which effective and accessible primary care exists to prevent, control, or manage these conditions.[46] ACSC hospitalizations are considered avoidable with adequate primary care, and are a reflection of unnecessarily use of health system resources.[47] Thus, ACSC hospitalizations are an indicator of health system performance. In the Canadian health system, hospitalizations for seven chronic ACSCs are routinely monitored, namely angina, asthma, congestive heart failure (CHF), COPD, diabetes and diabetic complications, epilepsy, and hypertension.[48] Hospitalizations for chronic ACSCs may more specifically indicate insufficient disease management.[49-54]
Multiple chronic diseases: A composite chronic disease endpoint reflects the shifting focus on disease specific prevention activities to recognition of targeting shared risk factors and a greater understanding on their collective impact on health. The four chronic diseases of interest are: cancer; cardiovascular disease consisting of chronic heart failure, myocardial infarction, and stroke including transient ischemic attack; chronic respiratory disease consisting of asthma and COPD, and diabetes. These major chronic diseases were chosen based on their high prevalence, overall impact on health and premature

mortality, and alignment with other chronic disease priority actions from the World Health Organization (WHO), the United Nations and the Public Health Agency of Canada (PHAC).[1,2,7] With the exception of cancer, all chronic diseases will be ascertained from health administrative data using chronic disease ascertainment algorithms (Appendix 1).

For each segmentation we will describe the population (number and proportion in each risk segment and the characteristics of those populations). To understand how the segments relate to our 6 outcomes we will use ridge regression or LASSO regression because they are particularly designed for a large set of predictors, where some predictors could have negligibly small effect.[55] For internal validation, we will calculate the 10-fold cross-validation prediction error of each model within the full database for all years. For external validation, we will subset the sample based on the data collection period.[56] We will estimate measures of discrimination using the C-statistic, mean absolute prediction error (MAPE), proportion of variation in the outcomes explained using R^2,[57] differences in outcome rates (mortality, chronic disease, high cost users) and costs within population segments. Overall the metrics will provide us with a robust assessment of how these approaches are performing in explaining variation and discrimination but also provide important information on how the population segments manifest and what important sub-groups may be missed for each approach. These will produce empirical results to support performance across a range of outcomes. We realize that there are many outcomes to summarize. Therefore, we will produce a range of tools to summarize the findings including heat maps, graphical approaches and summary statistics.

Objective 3 Present and test the different approaches to a decision-making team, responsible for population health management in an attributed population.

Our goal is to work closely with health system decision-makers at THP who are in the process of planning for how they will carry out population health management in the context of the new Ontario Health Teams. They will be implementing the tools generated from this research for an attributed population of approximately 800,000 individuals. We will build from our experience implementing decision analytic tools into practice. Specifically, the NPA led the development of the Population Health Planning Knowledge to Action Model (i.e., PHPKtoA Model – Appendix 2), designed specifically to support the development and use of population health risk tools in practice.[58] The PHPKtoA Model offers a systematic guide for researchers and health system partners to collaboratively develop meaningful population risk tools (i.e., the Tool Creation Path), and to facilitate use and application of the tools in diverse health settings (i.e., the Action Cycle Process). The appropriateness of the PHPKtoA Model and knowledge brokering (KB) strategy was demonstrated and evaluated in a similar integrated knowledge translation (IKT) initiative focused on collaboratively working with four health settings.[59 60] The Tool Creation Path – describes the analytic work that will take place in objectives 1 and 2. The actional cycle involves forming a partnership with THP and their partners, assessing needs and developing a common understanding to population segmentation goals. We will accomplish this with our knowledge translation specialist leading a series of in-person workshops. We will then proceed to the next phase which involves presenting the analytic outputs specifically for the population covered. Through a series of interactive workshops with our key decision-makers, we will build capacity with our health system partner to be able to understand and carry out the various stratification approaches. We have developed a series of data collection methods for users to participate in interviews to share experiences in using the tools and perceived outcomes. Evaluation strategies have been integrated into all stages of the Action Cycle process to enhance the attainment of stage-specific goals, and to strengthen and measure the impact of the IKT approach and tool use. Evaluation questions focused on process improvement, monitoring, and outcomes have been informed by intended

users and mapped to each stage in the model. An overview of our comprehensive evaluation strategy is presented in Appendix 4, and is organized according to the stages of the PHPKtoA Model.

Anticipated results and project outputs: (1) The development and application of new data-driven segmentation approaches; (2) Empirical comparison of existing and new segmentation approaches using data relevant for Canadian populations and health systems; (3) A validation of new and existing approaches against a range of health system and population health outcomes; (4) A formal partnership with a health system lead responsible for population health management as part of a major health system transformation; (5) The development of novel dissemination strategies to present the segmentation comparisons such that is accessible means for the decision-maker engagement. (6) Direct decision-support regarding analytic approaches to implement population health management.

Sample size considerations, feasibility, and timeline: An advantage of this study is that we are considering the entire population residing in Ontario during the study period (i.e. 14 million people), which means we are not selecting a sample over the study period. The survey data for some of the predictions analysis includes use of multiple cycles over several years, resulting in large cohorts (>200,000 individuals) available for analysis; therefore, we will be using the maximum sample size possible for this study. We have run several simulations to empirically show that the sample sizes in the proposed data sets will be adequate for generating measures of predictive accuracy. Given a 20% false positive rate, a sample size of 30,000 can detect 90% of the subjects who have any of the outcomes, while a sample size of 1,000 could detect about 85% of these cases. Simulations of calibration with different sample sizes were also run and demonstrate that they are typically bounded by 2% and stable across sample sizes. The detailed **timeline** of the study can be found in Appendix 1.

LIMITATIONS AND MITIGATION STRATEGIES

The cluster-based approaches we are applying are cutting edge and have significant advantages over more traditional approaches; however, as a result, they may have challenges with computational needs and convergence. Although we believe we have the computational resources for these approaches, alternative approaches, including centroid-based (such as k-means[61]) or connectivity-based approaches (such as PCA) will also be utilized, which are feasible to implement with our existing resources. For one of our approaches we will use population-level survey data that are routinely collected, which is advantageous as outputs can be updated frequently and remain relevant for different regions. However, these self-report data are potentially subject to misclassification error. We will conduct Quantitative Bias Analysis (QBA)[62] to estimate the impact of measurement error on our data. Finally, through out IKT strategy, we will evaluate the utility of the segmentation approaches in a large hospital network that is responsible for population health management for a community of 800,000 residents, which will inform generalizability of the segmentation approaches to other settings and areas for improvement.

RESEARCH, POLICY EXPERTISE AND CONTRIBUTIONS

We have assembled an expert team of health system leaders, methodologists and epidemiologists whose combined expertise is strongly positioned to carry out the proposed research. **Dr. Laura Rosella** is the Principal Investigator and Scientific Director of the Population Health Analytics Lab. She is an Associate Professor in the Dalla Lana School of Public Health at the University of Toronto (U of T), Faculty Member in the Vector Institute for Artificial Intelligence, and holds scientific appointments at ICES (Site Director for ICES U of T) and Public Health Ontario. She also holds a Canada Research Chair in Population Health Analytics. Her research interests are focused on developing analytic tools to support health decision-making. She has been awarded several national grants including one recently funded to advance the application of machine learning methods for studying multiple chronic diseases.

LR will oversee design and conduct of the overall project, specifically direct the application of the predictive models and co-lead the implementation of the new clustering approaches with MU. **Dr. Peter Austin** is a Senior Scientist at ICES and Professor at the Institute of Health Policy, Management and Evaluation (IHPME) at U of T. He has led the methodological evaluation of a range of expert-driven and predictive models related to segmentation. He has been recognized by Thomson Reuters and others. He is on the Highly Cited Researcher list (2015-2018), in many of these years being the only Canadian on this list. *He will provide critical expertise on the statistical approaches used to implement and evaluate the performance of the segmentation approaches. Specifically, he will oversee specific analysis for objective 1 and objective 2 related to the expert-driven methodologies.* **Dr. Robert Reid** is the Chief Scientist at the Institute for Better Health located within Trillium Health Partners. He is also a Professor at the U of T's IHPME. He also holds appointments as affiliate investigator at Group Health Research Institute, adjunct professor at the University of British Columbia's School of Population and Public Health, and affiliate associate professor at the University of Washington's School of Public Health and Community Medicine. Dr. Reid is best known for his work in developing and evaluating Group Health Cooperatives' pioneering patient-centred medical home in Seattle, Washington, which coordinated and streamlined team-based care to improve quality, patient experience and yield cost savings. A fellow of the American College of Preventive Medicine, he has lectured nationally and internationally on the organization, financing and delivery of care. Particularly relevant to this grant, he was the first Canadian investigator to apply the Johns Hopkins ACG system on Canadian data and lead the evaluation of its performance. *In addition to scientific expertise in applying and implementing segmentation approaches, he will also provide the critical link to Trillium Health Partners to operationalize objective 3 and the IKT approach*. **Dr. Walter Wodchis** is Professor at the IHPME at U of T, a Research Chair in Implementation and Evaluation Science, Institute for Better Health, Trillium Health Partners and an Adjunct Scientist at ICES. His main research interests are health economics and financing, health care policy evaluation, and long-term care. Dr. Wodchis is also the principal investigator for the Health System Performance Research Network where he leverages expertise in performance measurement in multiple domains of health system performance including clinical quality, financial management, patient safety and patient satisfaction to inform health system improvement in Ontario. He is an internationally recognized expert in using administrative data in longitudinal cost estimation analysis and leadership on health system evaluation, and he will provide senior leadership relevant to all aspects of this evaluation. *He will provide oversight on the use of the health services and costing data as well as co-supervise trainees working on the analysis using administrative data.* **Dr. Madeleine Udell** is Assistant Professor of Operations Research and Information Engineering and Richard and Sybil Smith Sesquicentennial Fellow at Cornell University. She studies optimization and machine learning for large scale data analysis and control, with applications in marketing, demographic modeling, medical informatics, engineering system design, and automated machine learning. Her paper "Why are big data matrices approximately low rank?" is currently the most read article in the SIMODS journal. Her research in optimization centers on detecting and exploiting novel structures in optimization problems, with a particular focus on convex and low rank problems, which are common to health data. She has developed a number of open source libraries for modeling and solving optimization problems, including Convex.jl, one of the top tools in the Julia language for technical computing. Her work on generalized low rank models (GLRMs) extends principal components analysis (PCA) to embed tabular data sets with heterogeneous (numerical, Boolean, categorical, and ordinal) types into a low dimensional space, providing a coherent framework for compressing, denoising, and imputing missing entries. Our collaboration is key to the success of this grant. *Specifically, she will provide specific expertise regarding the data-driven approaches and with the NPA plan to co-supervise a PhD student with the capabilities to carry out the novel data-driven analyses proposed in objective.*
The **research environment** is a combination of DLSPH at the University of Toronto with a rich

network of public health and health policy faculty and students and a link to Statistics Canada data through the RDC; ICES, which is the repository for the largest linked data holdings in Canada and holds all the expertise and data required for the proposed objectives. Finally, our formal letter of support from THP CEO as well as the participation of Dr, Reid on our Co-I team demonstrates the commitment and environment for objective 3.

KNOWLEDGE TRANSLATION

We have planned an integrated and interactive knowledge translation strategy that will capitalize on the reach and potential of this innovative work and integral partnership of our knowledge users and use their wide-ranging expertise to identify practical health system solutions. Briefly, the main components of our adapted European Science Advisory Network for Health (EuSANH)[63] includes: (i) Framing the issue (population segmentation from their perspective); (ii) Planning the process (we have devised a detailed work plan); (iii) Completing analyses and facilitating synthesis of findings (our team of foremost international experts on these methods will work collaboratively to oversee analyses and generate relevant findings that will be used by THP); (iv) Formulating the recommendations (generated by THP thus permitting consideration of a diverse range of segmentation options and potential risks/consequences); (v) Reviewing the findings/knowledge products (all findings/knowledge products will be reviewed by representatives from THP; external peer review will be sought); (vi) Disseminating the findings/knowledge products (findings will be shared at the organizational level and via open-access publications/accessible web-based platforms); and (vii) Assessing the impact (to support accountability, team performance will be assessed using the Partnership Self-Assessment Tool (PSAT),[64] which appraises partnership impact according to the Partnership Synergy Framework.[65] ***End of Grant Knowledge Translation:*** We will disseminate findings to other researchers, physicians, decision- and policy-makers, and key stakeholders through traditional mechanisms for research sharing such as conferences, peer-reviewed publications, and investigative reports. We will leverage the diverse institutes and universities that employ our investigators to share the results with academic communities. In addition, we will share our findings through the diverse professional networks represented by our research team (e.g., epidemiology, public health, medicine, health economics, biostatistics). The use of PSAT will assess effectiveness of the partnership with THP. PSAT is a commonly used and validated measure for assessing the effectiveness of partnerships, which is a critical dimension of our IKT approach. Findings from the PSAT will be shared with all team members and will be used to support enhancement of team functioning.

SIGNIFICANCE AND IMPACT OF THE RESEARCH

To summarize, this proposal will have meaningful impact in several ways. From a methodological perspective this study represents the largest and most comprehensive study in Canada testing and implementing both existing and new segmentation approaches. This will result in a significant contribution to rapidly growing scientific literature on data science approaches for health system management. Specifically, the application of the density-based clustering and generalized low rank models demonstrate the first application of these methods using population data and have the potential to translate the findings into action. Secondly, by comparing these approaches across a range of meaningful health system outcomes, our analysis will provide a direct way to measure the applied impact of these methods. Thirdly, for the first time in Canada, a practical guidance on how to implement population health management using data and different analytic approaches. This will result from our comprehensive evaluation and the third objective where we will gain a critically understanding of what approach is optimal and why through the lens of health decision-makers who are tasked with implementing these approaches. We feel we have put together the team and health system partners to achieve important impact with this proposal.

References

1. Sox HC. Resolving the Tension Between Population Health and Individual Health Care. *Jama-Journal of the American Medical Association* 2013;310(18):1933-34. doi: 10.1001/jama.2013.281998
2. Vuik SI, Mayer, E.K., Darzi, A. . Patient Segmentation Analysis Offers Significant Benefits For Integrated Care And Support. *Health Affairs* 2016;35(5):769-75. doi: 10.1377/hlthaff.2015.1311
3. Rosen AK, Reid R, Broemeling A-M, et al. Applying a risk-adjustment framework to primary care: can we improve on existing measures? *The Annals of Family Medicine* 2003;1(1):44-51.
4. Kindig D, Stoddart G. What is population health? *American Journal of Public Health* 2003;93(3):380-83. doi: 10.2105/ajph.93.3.380
5. Canadian Institute for Health Information. CIHI's Population Grouping Methodology 1.1: Methodology Report. . Ottawa, Ontario, 2017.
6. Yan JL, Linn KA, Powers BW, et al. Applying Machine Learning Algorithms to Segment High-Cost Patient Populations. *Journal of General Internal Medicine* 2019;34(2):211-17. doi: 10.1007/s11606-018-4760-8
7. Madeleine U, Corinne H, Reza Z, et al. Generalized Low Rank Models: now 2016.
8. Discovering patient phenotypes using generalized low rank models. Biocomputing 2016: Proceedings of the Pacific Symposium; 2016. World Scientific.
9. Udell M. LowRankModels.jl 2019.
10. Graph-regularized generalized low-rank models. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops; 2017.
11. Manuel DG, Rosella L, Hennessy D, et al. Predictive risk algorithms in a population setting: An overview. *Journal of Epidemiology and Community Health* 2012;66:859-65.
12. Stone C, Rosella L, Goel V. Population health perspective on high users of health care: Role of family physicians. Canadian Family Physician, 2014:781-83.
13. Rosella LC, Kornas K, Yao Z, et al. Predicting High Health Care Resource Utilization in a Single-payer Public Health Care System: Development and Validation of the High Resource User Population Risk Tool. *Medical Care* 2018;56(10):E61-E69. doi: 10.1097/mlr.0000000000000837
14. Manuel DG, Rosella LC, Tuna M, et al. Effectiveness of Community-Wide and Individual High-Risk Strategies to Prevent Diabetes: A Modelling Study. *Plos One* 2013;8(1) doi: 10.1371/journal.pone.0052963
15. Rosella LC, Lebenbaum M, Li Y, et al. Risk distribution and its influence on the population targets for diabetes prevention. *Preventive Medicine* 2014;58:17-21. doi: 10.1016/j.ypmed.2013.10.007
16. Wennberg D SM, Darin B, et al. . Combined predictive model: final report and technical documentation,. Cambridge, UK, 2006.
17. Chenore T, Pereira Gray D, Forrer J, et al. Emergency hospital admissions for the elderly: insights from the Devon Predictive Model. *Journal of Public Health* 2013;35(4):616-23.
18. Rosella LC, Kornas, K. Putting a Population Health Lens to Multimorbidity in Ontario. *Healthcare Quarterly* 2018;21(3):8-11.
19. Rosella LC, Fitzpatrick T, Wodchis WP, et al. High-cost health care users in Ontario, Canada: demographic, socio-economic, and health status characteristics. *Bmc Health Services Research* 2014;14 doi: 10.1186/s12913-014-0532-2
20. Jeffery AD, Hewner S, Pruinelli L, et al. Risk prediction and segmentation models used in the United States for assessing risk in whole populations: a critical literature review with

implications for nurses' role in population health management. *JAMIA Open* 2019;2(1):205-14. doi: 10.1093/jamiaopen/ooy053
21. Choi Y, Chiu CY-I, Sontag D. Learning Low-Dimensional Representations of Medical Concepts. *AMIA Jt Summits Transl Sci Proc* 2016;2016:41-50.
22. Peretti-Watel P, Spire B, Lert F, et al. Drug use patterns and adherence to treatment among HIV-positive patients: evidence from a large sample of French outpatients (ANRS-EN12-VESPA 2003). *Drug and Alcohol Dependence* 2006;82:S71-S79. doi: 10.1016/s0376-8716(06)80012-8
23. Rosella LC, Fitzpatrick T, Wodchis WP, et al. High-cost health care users in Ontario, Canada: demographic, socio-economic, and health status characteristics. *BMC health services research* 2014;14:532. doi: 10.1186/s12913-014-0532-2 [published Online First: 2014/11/02]
24. Bilandzic A, Rosella L. The cost of diabetes in Canada over 10 years: applying attributable health care costs to a diabetes incidence prediction model. *Health promotion and chronic disease prevention in Canada : research, policy and practice* 2017;37(2):49-53. [published Online First: 2017/03/09]
25. Rosella LC, Lebenbaum M, Fitzpatrick T, et al. Impact of diabetes on healthcare costs in a population-based cohort: a cost analysis. *Diabetic medicine : a journal of the British Diabetic Association* 2016;33(3):395-403. doi: 10.1111/dme.12858 [published Online First: 2015/07/24]
26. Canadian Institute for Health Information.
27. Statistics Canada. Canadian Community Health Survey 2003: User Guide for the Public Use Microdata File. Ottawa, ON, Canada: Statistics Canada; 2005. (Catalogue no. 82M0013GPE).
28. Ontario Agency for 761 Health Protection and Promotion Act. Q1 762, 2007.
29. Care OMoHaL-T. About the Ministry 2013 [updated 4/3/2013. Available from: http://www.health.gov.on.ca/en/common/ministry/default.aspx accessed 8/17/2014 2014.
30. Chiu M, Lebenbaum M, Lam K, et al. Describing the linkages of the immigration, refugees and citizenship Canada permanent resident data and vital statistics death registry to Ontario's administrative health database. *BMC medical informatics and decision making* 2016;16(1):135.
31. Johns Hopkins ACG Case-Mix Adjustment System [program], 2016.
32. Austin PC, Stanbrook MB, Anderson GM, et al. Comparative ability of comorbidity classification methods for administrative data to predict outcomes in patients with chronic obstructive pulmonary disease. *Annals of epidemiology* 2012;22(12):881-87. doi: 10.1016/j.annepidem.2012.09.011 [published Online First: 10/31]
33. Austin PC, van Walraven C, Wodchis WP, et al. Using the Johns Hopkins Aggregated Diagnosis Groups (ADGs) to predict mortality in a general adult population cohort in Ontario, Canada. *Medical care* 2011;49(10):932-39. doi: 10.1097/MLR.0b013e318215d5e2
34. Austin PC, Walraven Cv. The mortality risk score and the ADG score: two points-based scoring systems for the Johns Hopkins aggregated diagnosis groups to predict mortality in a general adult population cohort in Ontario, Canada. *Medical care* 2011;49(10):940-47. doi: 10.1097/MLR.0b013e318229360e
35. Reid RJ, MacWilliam L, Roos N, et al. Measuring morbidity in populations: performance of the Johns Hopkins adjusted clinical group (ACG) case-mix adjustment system in Manitoba. Winnipeg, MB: Manitoba Centre for Health Policy and Evaluation 1999.
36. Reid RJ, MacWilliam L, Verhulst L, et al. Performance of the ACG case-mix system in two Canadian provinces. *Medical care* 2001:86-99.
37. Reid RJ, Roos NP, MacWilliam L, et al. Assessing population health care need using a claims-based ACG morbidity measure: a validation analysis in the Province of Manitoba. *Health Serv Res* 2002;37(5):1345-64.
38. OPTICS: ordering points to identify the clustering structure. ACM Sigmod record; 1999. ACM.
39. A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd; 1996.

40. O'Neill MK, K; Buajitti, E; Bornbaum, C; Wodchis, W; Rosella, LC Predicting high healthcare resource utilization in Ontario, 2013/14-2018/19: Characterizing high resource users in Public Health Units. Toronto, Ontario, 2019.
41. Bootstrap variance estimation for predicted individual and population-average risks. Proceedings of the Survey Research Methods Section Alexandria VA: American Statistical Association; 2008.
42. Bootstrap variance estimation for the national population health survey. American Statistical Association, Proceedings of the Survey Research Methods Section; 1999. Citeseer.
43. Wodchis WP, Austin PC, Henry DA. A 3-year study of high-cost users of health care. *Canadian Medical Association Journal* 2016
44. Pefoyo AJK, Bronskill SE, Gruneir A, et al. The increasing burden and complexity of multimorbidity. *BMC Public Health* 2015;15(1):415.
45. Wodchis W, Bushmeneva K, Nikitovic M, et al. Guidelines on Person-Level Costing Using Administrative Databases in Ontario. Toronto 2013.
46. Billings J, Zeitel L, Lukomnik J, et al. Impact of socioeconomic status on hospital use in New York City. *Health Aff (Millwood)* 1993;12(1):162-73. doi: 10.1377/hlthaff.12.1.162
47. Avoidable Hospitalization Advisory Panel. Enhancing the Continuum of Care: Report of the Avoidable Hospitalization Advisory Panel. Ottawa, ON, 2011.
48. Canadian Institute for Health Information. Ambulatory Care Sensitive Conditions Ottawa, ON: CIHI; 2018 [accessed March 11 2019.
49. Brown AD, Goldacre MJ, Hicks N, et al. Hospitalization for ambulatory care-sensitive conditions: a method for comparative access and quality studies using routinely collected statistics. *Can J Public Health* 2001;92(2):155-9.
50. Busby J, Purdy S, Hollingworth W. How do population, general practice and hospital factors influence ambulatory care sensitive admissions: a cross sectional study. *BMC Fam Pract* 2017;18(1):67. doi: 10.1186/s12875-017-0638-9
51. Canadian Institute for Health Information. Disparities in primary health care experiences among Canadians with ambulatory care sensitive conditions. Ottawa, ON, 2012.
52. Daniels LM, Sorita A, Kashiwagi DT, et al. Characterizing Potentially Preventable Admissions: A Mixed Methods Study of Rates, Associated Factors, Outcomes, and Physician Decision-Making. *J Gen Intern Med* 2018;33(5):737-44. doi: 10.1007/s11606-017-4285-6
53. Gilotra NA, Shpigel A, Okwuosa IS, et al. Patients Commonly Believe Their Heart Failure Hospitalizations Are Preventable and Identify Worsening Heart Failure, Nonadherence, and a Knowledge Gap as Reasons for Admission. *J Card Fail* 2017;23(3):252-56. doi: 10.1016/j.cardfail.2016.09.024
54. Porter J, Herring J, Lacroix J, et al. Avoidable admissions and repeat admissions: what do they tell us? *Healthc Q* 2007;10(1):26-8.
55. Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2011;73(3):273-82.
56. Mukherjee B, Ou H-T, Wang F, et al. A new comorbidity index: the health-related quality of life comorbidity index. *Journal of Clinical Epidemiology* 2011;64(3):309-19. doi: 10.1016/j.jclinepi.2010.01.025
57. Chang H-Y, Weiner JP. An in-depth assessment of a diagnosis-based risk adjustment model based on national health insurance claims: the application of the Johns Hopkins Adjusted Clinical Group case-mix system in Taiwan. *BMC Med* 2010;8:7-7. doi: 10.1186/1741-7015-8-7
58. Peirson L, Rosella L. Navigating Knowledge to Action: A conceptual map for facilitating translation of population health risk planning tools into practice. *Continuing Education in the Health Professions* 2015;35(2):139-47.

59. Rosella L, Peirson L, Bornbaum C, et al. Supporting collaborative use of the Diabetes Population Risk Tool (DPoRT) in health-related practice: a multiple case study research protocol. *Implementation Science* 2014;9(1):35. doi: 10.1186/1748-5908-9-35
60. Rosella L, Bornbaum, C., Kornas, K., Lebenbaum, M, Loeppky, C., Gardner, C., Mowat, D. Evaluating the Process and Outcomes of a Knowledge Translation Approach to Supporting Use of the Diabetes Population Risk Tool (DPoRT) in Public Health Practice *Canadian Journal of Program Evaluation* 2018;33(1):21-48.
61. Vuik SI, Mayer E, Darzi A. A quantitative evidence base for population health: applying utilization-based cluster analysis to segment a patient population. *Population Health Metrics* 2016;14 doi: 10.1186/s12963-016-0115-z
62. Lash TL, Fox MP, MacLehose RF, et al. Good practices for quantitative bias analysis. *International journal of epidemiology* 2014;43(6):1969-85.
63. Sarría-Santamera A, Schoten EJ, Coenen TMM, et al. A framework for scientific advice on health: EuSANH's principles and guidelines. *Health Research Policy and Systems* 2013;11(1):6. doi: 10.1186/1478-4505-11-6
64. Cramm JM, Strating MM, Nieboer AP. Development and validation of a short version of the Partnership Self-Assessment Tool (PSAT) among professionals in Dutch disease-management partnerships. *BMC research notes* 2011;4(1):224.
65. Lasker RD, Weiss ES, Miller R. Partnership synergy: a practical framework for studying and strengthening the collaborative advantage. *Milbank Q* 2001;79(2):179-205, iii-iv. [published Online First: 2001/07/07]