HIGH-COST HEALTH CARE USERS IN SASKATCHEWAN:
A POPULATION HEALTH PERSPECTIVE

BY

MAUREEN F. ANDERSON

A Thesis
Submitted to the Graduate Faculty
in Partial Fulfilment of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

Department of Health Management
Atlantic Veterinary College
University of Prince Edward Island

©MAY 2019 M.F. Anderson.
In consideration of my University making my thesis/dissertation available to interested persons, I,

Maureen Anderson

hereby grant a non-exclusive, for the full term of copyright protection, license to my University,

The University of Prince Edward Island:

(a) to archive, preserve, produce, reproduce, publish, communicate, convert into any format, and to make available in print or online by telecommunication to the public for non-commercial purposes;
(b) to sub-license to Library and Archives Canada any of the acts mentioned in paragraph (a).

I undertake to submit my thesis/dissertation, through my University, to Library and Archives Canada. Any abstract submitted with the thesis/dissertation will be considered to form part of the thesis/dissertation.

I represent that my thesis/dissertation is my original work, does not infringe any rights of others, including privacy rights, and that I have the right to make the grant conferred by this non-exclusive license.

If third party copyrighted material was included in my thesis/dissertation for which, under the terms of the Copyright Act, written permission from the copyright owners is required I have obtained such permission from the copyright owners to do the acts mentioned in paragraph (a) above for the full term of copyright protection

I retain copyright ownership and moral rights in my thesis/dissertation, and may deal with the copyright in my thesis/dissertation, in any way consistent with rights granted by me to my University in this non-exclusive license.

I further promise to inform any person to whom I may hereafter assign or license my copyright in my thesis/dissertation of the rights granted by me to my University in this non-exclusive license.

<table>
<thead>
<tr>
<th>Signature</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Date</th>
</tr>
</thead>
</table>

July 18, 2019
University of Prince Edward Island

Faculty of Veterinary Medicine

Charlottetown

CERTIFICATION OF THESIS WORK

I, the undersigned, certify that Maureen Frances Anderson, BSc, MSc, candidate for the degree of Doctor of Philosophy (PhD) has presented his/her thesis with the following title:

High-cost users of health care in Saskatchewan: A population health perspective

that the thesis is acceptable in form and content, and that a satisfactory knowledge of the field covered by the thesis was demonstrated by the candidate through an oral examination held on June 6th, 2019.

Examiners

Dr. (External) ______________________

Dr. (Chair) ______________________

Dr. (Supervisor) ______________________

Dr. ______________________

Dr. ______________________
ABSTRACT

Background: A small proportion of the population consumes the majority of health-care resources. High-cost user research is complicated by heterogeneous populations, a natural tendency for a regression to the mean and episodic versus persistent spending patterns. Using two separate study cohorts, 1) a clinical sub-group known to be high-cost – mental health and addiction clients, 2) a general provincial population grouped into separate clinically meaningful sub-groups, this thesis seeks to understand the high-cost health care user population in the province of Saskatchewan, Canada.

Methods: The first two quantitative studies focus on a specific disease sub-population known to be high-cost – mental health and addiction clients. First, a retrospective cohort study explores risk factors associated with high-cost use with a particular focus on individuals who are persistently high-cost year after year. The second study aims to predict individuals at risk of both episodic and persistent high-cost use in the future. Predictive models using Classification and Regression Tree (CART) methods were constructed. The last study takes an overall population segmentation approach to understanding high-cost use. Using the recently developed Canadian Institute for Health Information Population Grouping methodology, individuals were assigned to one of 16 mutually exclusive and clinically distinct health profile groups. Following univariate and bivariate analyses, logistic regression models were constructed for each of the costliest health profile groups to explore risk factors associated with high-cost health care use.

Results: Study 1: Persistent high-cost mental health and addiction clients comprised a small proportion of the study cohort (n = 6,455; 5%) but accounted for 35% of total costs. Exploratory models of mental health and addiction high-cost patients found increased risk of persistent high-cost use with hospitalization(s), unstable housing, severity of the diagnosis (schizophrenia versus others), and multiple comorbidities. Good connection to a primary care provider was protective of high-cost use, particularly when individuals had multiple mental health conditions.

Study 2: The most important variables for predicting one-year and persistent high-cost use were delineated visually in CART diagrams. My models had reasonable calibration and validation and take advantage of health care utilization and demographic information readily available in Canadian provincial administrative health care databases. The visual nature of the CART method assists to make complex data readily understandable to policy and decision-makers.

Study 3: A provincial cohort (n = 1,175,147) was identified for study. High-cost users consumed 41% of total health care resources. The costliest health profile groups were ‘long-term care’, ‘palliative’, ‘major acute’, ‘major chronic’, ‘major cancer’, ‘major newborn’, ‘major mental health’ and ‘moderate chronic’. Both ‘major acute’ and ‘major cancer’ health profile groups were largely explained by measures of health care utilization and multi-morbidity. In the remaining costliest health profile groups modelled, ‘major chronic’, ‘moderate chronic’, ‘major newborn’
and ‘other mental health’, a measure of socio-economic status, low neighbourhood income, was statistically significantly associated with high-cost use.

In each exploratory study, when controlling for a variety of factors including demographics, health care utilization and health status, baseline measures of socio-economic status – unstable housing and low neighbourhood income – were found to be statistically significantly associated with high-cost use.

**Conclusion:** Interventions aimed at improving population health and reducing the health care costs associated with a ‘high-cost user’ population should consider segmenting the population into relevant homogenous sub-groups, defining high-cost use within sub-groups, and, exploring risk factors associated with high-cost use within each subgroup. Primary health care system transformation efforts could include a ‘high-cost user’ component. Given that having a good connection to a primary care provider was found to be protective of high-cost use, it is suggested that primary care providers, and, high-cost patients themselves design interventions aimed to reduce costs and improve population health. Lastly, socio-economic status must be considered in exploratory and predictive modelling of high-cost health care use; policy efforts to address socio-economic status in a high-cost population may result in health care system cost savings, but more importantly, improved population health.
ACKNOWLEDGMENTS

This dissertation is the culmination of over 6 years of hard work and dedication – none of which I could have done alone.

First, I sincerely thank Drs. Carolyn Sanford and Heather Morrison, Prince Edward Island Department of Health and Wellness for their support, encouragement and practical advice; without which this journey would never have begun. To Robin Buckland, Samantha Wilson-Clark and Dr. Tig Shafto at the Public Health Agency of Canada (PHAC) for bravely stating, in many a written memo, that PHAC should ‘put their money where their mouth is’ when it came to increasing epidemiological capacity in Canada and advocating for my PhD-level training. To Dr. Carol McClure – the best office-mate (and advice giver, supportive sense of humour, baby shower thrower and stats brain) around.

To the Saskatchewan “Hot-spotting CIHR grant” research team, Drs. Gary Teare, Cory Neudorf, Nazeem Muhajarine, Marilyn Baetz, Jacqueline Quail and Ms. Margaret Baker – thank you for permitting me to join your team and piggyback my dissertation onto the overall grant. Thank you to all of the analysts and privacy specialists, and most especially Joanne Franko, who worked so hard to enable the linkage of disparate databases that had never before come together in Saskatchewan. To my Ontario collaborators, Drs. Walter Wodchis, Claire de Oliveria and Laura Rosella – I have learned much from you.

To my co-supervisors, Drs. Crawford Revie and Cory Neudorf, my gratitude for your practical, calm, brilliant mentorship, guidance and unwavering support cannot be overstated. To Dr. Henrik Stryhn – I have never, and likely will never, meet a better biostatistics professor; I cannot thank you enough for all that you have taught me. To Dr. J McClure – despite the topic being wildly out your content area you were always there to lend a helping hand and support this, oftentimes unwieldy, project coming to fruition; I couldn’t have asked for a better committee chair. To Dr. David Buckeridge – thank you for agreeing to be part of this committee and sharing your valuable time, content knowledge, strokes of genius and support.

To my epi colleagues I am lucky enough to also call friends: Drs. Shauna Richards, Lisa Jensen, Tammy Stuart-Chester, Kate Zinzser and Jacqueline Quail – there were days your encouragement and support were all that kept me going.

To the goddesses – Leanne Kadyschuk, Virginia Scissors, Vanessa Hyggen and Laelia LeFeuvre – life is better with your support and love in it.

And most of all – thank you to my husband Aaron for everything. Words, quite literally, cannot express all you have done to make this possible. Love you so very much.
DEDICATION

To my daughters – Eleanor and Charlotte

I love you both a million times more than the feeling I anticipate that will come over me when this PhD journey is complete. That is a lot, my loves.
# TABLE OF CONTENTS

**ABSTRACT** .......................................................................................................................... iv  
**ACKNOWLEDGMENTS** ................................................................................................. vi  
**DEDICATION** .................................................................................................................. vii  
**TABLE OF CONTENTS** ................................................................................................. viii  
**LIST OF TABLES** ........................................................................................................... xi  
**LIST OF FIGURES** ........................................................................................................... xii  
**LIST OF ABBREVIATIONS** ........................................................................................... xiii  
1. **INTRODUCTION** .......................................................................................................... 1  
   1.1 Rationale for study .................................................................................................... 1  
   1.2 Objectives ................................................................................................................ 2  
      1.2.1 *Theoretical framework* .................................................................................. 3  
   1.3 Literature review ...................................................................................................... 5  
      1.3.1 Health care costing in Canada ....................................................................... 5  
      1.3.2 Definition of ‘high-cost’ health care use ...................................................... 11  
      1.3.3 Statistical modelling of health care costs ..................................................... 13  
      1.3.4 High-cost health care literature .................................................................. 15  
   1.4 Organization of thesis ............................................................................................... 28  
   1.5 References ............................................................................................................... 33  
2. **EFFECT OF SOCIO-DEMOGRAPHIC FACTORS ON HIGH-COST MENTAL HEALTH AND ADDICTION HIGH-COST USE: A RETROSPECTIVE, POPULATION-BASED STUDY IN SASKATCHEWAN** ......................................................................................................................... 47  
   2.1 Abstract ..................................................................................................................... 47  
   2.2 Introduction ............................................................................................................... 48  
   2.3 Methods ..................................................................................................................... 50  
   2.4 Results ......................................................................................................................... 56  
   2.5 Discussion .................................................................................................................... 63  
   2.6 References .................................................................................................................. 68  
   2.7 Supplemental material ............................................................................................... 76  
      2.7.1 Cohort creation ................................................................................................. 76  
      2.7.2 Health care costing data ................................................................................. 77  
      2.7.3 Other databases .............................................................................................. 77
3. USING CLASSIFICATION AND REGRESSION TREE ANALYSIS TO PREDICT FUTURE HIGH-COST HEALTH CARE USE: RESULTS FROM A RETROSPECTIVE POPULATION-BASED COHORT STUDY OF MENTAL HEALTH AND ADDICTION CLIENTS IN SASKATOON, SASKATCHEWAN .......................................................... 90

3.1 Abstract ...................................................................................................................... 90
3.2 Introduction .................................................................................................................. 91
3.3 Methods ...................................................................................................................... 95
3.4 Results ......................................................................................................................... 100
3.5 Discussion ................................................................................................................... 109
  3.5.1 Summary of findings ................................................................................................. 109
  3.5.2 Predictive models in practice .................................................................................... 111
  3.5.3 Interventions for high-cost healthcare use ................................................................. 113
  3.5.4 Limitations ............................................................................................................... 113
3.6 Conclusion ................................................................................................................... 115
3.7 References .................................................................................................................. 116
3.8 Supplemental material ............................................................................................... 121
  3.8.1 Cohort and variable creation .................................................................................... 121
  3.8.2 Statistical modeling ................................................................................................. 122
  3.8.3 Co-author affiliations and contributions ..................................................................... 124

4. DEFINING ‘ACTIONABLE’ HIGH-COST HEALTH CARE USE: RESULTS USING THE CANADIAN INSTITUTE FOR HEALTH INFORMATION POPULATION GROUPING METHODOLOGY .................................................................. 126

4.1 Abstract ....................................................................................................................... 126
4.2 Introduction ................................................................................................................... 127
  4.2.1 Theoretical framework ............................................................................................. 129
4.3 Methods ......................................................................................................................... 131
4.4 Results ......................................................................................................................... 139
LIST OF TABLES

Table 2.1: Mental health and addiction study cohort, Saskatoon Health Region, Saskatchewan (n = 129,932) ...........................................................................................................58
Table 2.2: Average annual cost and health care utilization by mental health diagnosis, Saskatoon Health Region, Fiscal years 2009–2015 (n = 129,932) .............................................................................................59
Table 2.3: Regression modelling, binary distribution, probit link, comparing persistent high-cost use, mental health and addiction cohort, excluding long-term care residents, Saskatoon Health Region, FY2009–2015 (n = 124,497) ...........................................................................................................62
Table 3.1 Mental health and addiction high-cost cases, Saskatoon, Saskatchewan, demographics and health care utilization by cost outcome (Study cohort n = 124,497) .................................................................104
Table 3.2 Confusion matrix, validation data, predicting one-time high-cost use, mental health and addiction clients, Saskatoon, Saskatchewan (n = 49, 875) ...........................................................................................................108
Table 3.3 Confusion matrix, validation data, predicting persistent high-cost use, mental health and addictions clients, Saskatoon, Saskatchewan (n = 49, 875) ...........................................................................................................108
Table 3.4 Performance measures, high-cost health care use predictive models .................................................115
Table 4.1 Descriptive epidemiology, demographics and health care utilization variables, by cost category, study population, April 1 2015 – March 31 2017 (n = 1,175,147) .........................................................................................144
Table 4.2: Mean health care costs, by health profile category and high-cost use, Saskatchewan, excluding long-term care residents, April 1 2015 to March 31 2017 (n = 1,175,147) ..............................................................146
Table 4.3 High-cost use within major chronic health profile group, odds ratios (Total n = 37,803; High-cost users n = 3,781) ...........................................................................................................147
Table 4.4 High-cost users within moderate chronic health profile group, odds ratios (Total n = 112,776; High-cost users n = 11,277) ...........................................................................................................147
Table 4.5 High-cost use within major newborn health profile group, odds ratio (Total n = 2,678; High cost users n = 267) .........................................................................................................................148
Table 4.6 High-cost users within other mental health profile group (Total n = 62,381; High-cost users n = 6,241) ..............................................................................................................................148
Table 4.7 High-cost users within major acute disease health profile group (Total n=34,733; High cost users n=3,473) .........................................................................................................................148
Table 4.8 High-cost users within major cancer health profile group (Total n=8,711; High-cost users n=872) ...........................................................................................................................................149
LIST OF FIGURES

Figure 2.1: Study cohort inclusion/exclusion criteria ................................................................. 57
Figure 2.2: Mental health and addiction study cohort by total yearly cost, Saskatoon Health Region, Saskatchewan, Fiscal years 2009–2015 (n = 129,932) .................................................................................................................. 60
Figure 3.1: Study inclusion/exclusion criteria ........................................................................ 103
Figure 3.2: Classification and regression tree analyses to predict high-cost use in one year, mental health and addiction cohort, Saskatoon, Saskatchewan (n = 124,497; training = 74,622; validation = 49,875). 106
Figure 3.3: Classification and regression tree analyses to predict persistent high-cost use, mental health and addiction cohort, Saskatoon, Saskatchewan (n = 124,497; training = 74,622; validation = 49,875). 107
Figure 4.1: Study cohort inclusion/exclusion criteria ................................................................. 143
Figure 4.2: Clinical profiles before and after applying clinical overrides, illustration, Population Grouping Methodology, CIHI, April 2017 ....................................................................................................................... 164
Figure 4.3: Health profile groups by severity, CIHI Population Grouping methodology ........ 166
LIST OF ABBREVIATIONS

ACE  Adverse Childhood Experiences
AIC  Akaike information criterion
ALC  Alternate Level of Care
AUC  Area under the curve
BMJ  British Medical Journal
CAD  Coronary artery disease
CAMH  Centre for Addiction and Mental Health
CART  Classification and Regression Tree
CCDSS  Canadian Chronic Disease Surveillance System
CHF  Congestive heart failure
CI  Confidence interval
CIHI  Canadian Institute for Health Information
CIHR  Canadian Institute for Health Research
CMAJ  Canadian Medical Association Journal
CMG  Case Mix Groups™
COPD  Chronic Obstructive Pulmonary Disease
CT  Connecticut
DAD  Discharge Abstract Database
DOI  Digital object identifier
DR.  Doctor
DRG  Diagnosis-related groups
DX  Disease
ED  Emergency department
FIG.  Figure
FL  Florida
FY  Fiscal year
HQC  Health Quality Council (Saskatchewan)
HMO  Health Management Organization
HX  History
IBD  Inflammatory Bowel Disease
ICD  International Classification of Diseases
ICD-9  International Classification of Diseases version 9
ICD-10-CA  International Classification of Diseases version 10, Canadian
ID  Identification
ICES  Institute for Clinical Evaluative Sciences
INSPQ  Institut national de santé publique du Québec
IT  Information technology
JAMA  Journal of the American Medical Association
LASSO  Least absolute shrinkage and selection operator
LOS  Length of stay in hospital
LTC  Long-term care
MCHP  Manitoba Centre for Health Policy
MHA  Mental health and addiction(s)
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIS</td>
<td>Management information system (finance)</td>
</tr>
<tr>
<td>NACRS</td>
<td>National Ambulatory Care Reporting System</td>
</tr>
<tr>
<td>NCBI</td>
<td>National Center for Biotechnology Information</td>
</tr>
<tr>
<td>NC</td>
<td>North Carolina</td>
</tr>
<tr>
<td>NHS</td>
<td>National Health System (UK)</td>
</tr>
<tr>
<td>NJ</td>
<td>New Jersey</td>
</tr>
<tr>
<td>OECD</td>
<td>Organization for Economic Cooperation and Development</td>
</tr>
<tr>
<td>ON</td>
<td>Ontario</td>
</tr>
<tr>
<td>OR</td>
<td>Odds ratio</td>
</tr>
<tr>
<td>PHAC</td>
<td>Public Health Agency of Canada</td>
</tr>
<tr>
<td>PHRS</td>
<td>Person Health Registry System</td>
</tr>
<tr>
<td>PHD</td>
<td>Doctorate of Philosophy</td>
</tr>
<tr>
<td>POP GROUPEr</td>
<td>Population grouping methodology</td>
</tr>
<tr>
<td>RAI-LTC</td>
<td>Resident assessment information for long term care</td>
</tr>
<tr>
<td>RAI-MDS</td>
<td>Resident assessment information – minimum data set</td>
</tr>
<tr>
<td>RCMP</td>
<td>Royal Canadian Mounted Police</td>
</tr>
<tr>
<td>RIW</td>
<td>Resource Intensity Weight</td>
</tr>
<tr>
<td>SAS</td>
<td>Statistical Analysis Software</td>
</tr>
<tr>
<td>SCI</td>
<td>Shared Client Index</td>
</tr>
<tr>
<td>SHR</td>
<td>Saskatoon Health Region</td>
</tr>
<tr>
<td>SK</td>
<td>Saskatchewan</td>
</tr>
<tr>
<td>SD</td>
<td>Standard deviation</td>
</tr>
<tr>
<td>SES</td>
<td>Socio-economic status</td>
</tr>
<tr>
<td>SQL</td>
<td>Structured query language</td>
</tr>
<tr>
<td>UPC</td>
<td>Usual provider continuity score</td>
</tr>
<tr>
<td>US</td>
<td>United States</td>
</tr>
<tr>
<td>UK</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>VA</td>
<td>Veteran’s Affairs (US)</td>
</tr>
<tr>
<td>WHO</td>
<td>World Health Organization</td>
</tr>
<tr>
<td>YR</td>
<td>Year</td>
</tr>
</tbody>
</table>
1. INTRODUCTION

1.1 Rationale for study

Health care systems in Canada and around the world are challenged with increased costs. Health care expenditure in nearly every developed country meets or exceeds 10% of gross domestic product (World Health Organization, 2015). Evidence from Canada and abroad consistently demonstrates that a small proportion of the population (~5-10%) accounts for the majority (60-80%) of total health care spending (Densen, Shapiro, & Einhorn, 1959; Roos, Burchill, & Carriere, 2003; Roos, Shapiro, & Tate, 1989; Stuart & Weinrich, 1998); these individuals are commonly referred to as ‘high-cost users’.

High-cost users are a heterogeneous population. In general, previous studies have found high-cost health care use to be associated with: 1) complex, multiple chronic conditions (multimorbidity), 2) catastrophic illness (for example, motor vehicle accident or major cancer), 3) high costs at end-of-life, 4) mental health and addictions, 5) institutional living, and 6) various indicators of lower socio-economic status (for example, food insecurity and poverty) (see Section 1.3.4). Typically, in examining factors associated with high-cost use, studies focus on one specific population of high-cost users, such as one geographic area, or, more commonly, a specific diagnostic grouping. Inpatient stays in acute care facilities (hospitalizations) are described as the costliest source of health care spending, but, prescription drug costs have also been shown to be major cost drivers (Chang et al., 2016; Kaufman, Ali, DeFiglio, Craig, & Brenner, 2014; Wammes, van der Wees, Tanke, Westert, & Jeurissen, 2018).

At the individual-level, high-cost health care use can persist over time, or, it can be a transient state (Chang et al., 2016; Hwang, LaClair, Camacho, & Paz, 2015; Ku, Chiou, & Liu, 2015;
Robst, 2015). Further complicating research into understanding high-cost use is a phenomenon known as ‘regression to the mean’. There is a natural tendency for high health care spending at the individual-level to return to average levels over time. Regression to the mean, in its broadest sense, can best be thought of as an empirical observation that appears extreme at first, but, will eventually swing back to normal. In the context of individual health care spending ‘regression to the mean’ typically occurs due to recovery from illness (Chakravarty & Cantor, 2016).

Despite these known challenges and limitations, there is, particularly amongst health care insurers, great interest in better understanding the small proportion of the population that consumes the majority of health care resources.

Understanding the demographics, clinical profiles, health care utilization patterns and predictors associated with high-cost health care use is important for identifying opportunities for upstream prevention; to determine the best value for money spent (value-based health care) and, most importantly, to improve the health of the population.

1.2 Objectives

The goal of this research was to use linkable, population-based, administrative health data and person-level health care costing data in Saskatchewan, Canada to identify, describe and predict high-cost health care users. Specific study objectives were:

1. To identify and describe individuals who: 1) are never high-cost, 2) are sometimes high-cost, and, 3) are persistently high-cost, among a cohort of mental health and addiction clients in Saskatoon, Saskatchewan.
a. A subsidiary aim was, among a cohort of mental health and addiction clients, to understand risk factors associated with persistent high-cost health care use, including measures of socio-economic status at the individual-level (Chapter 2).

2. To predict those at risk of high-cost health care use in the future among a cohort of mental health and addiction clients in Saskatoon, Saskatchewan.
   a. The aim was to develop predictive models for both 1) episodic high-cost use, and, 2) persistent high-cost use (Chapter 3).

3. To identify and describe high-cost health care use at the population level in Saskatchewan, Canada.
   a. A subsidiary aim was, from the perspective of a provincial health care insurer, to quantify risk factors associated with high-cost health care use within mutually exclusive diagnostic groups (Chapter 4).

1.2.1 Theoretical framework

As the research reported in this thesis is based on health care utilization data, it is important to understand the theoretical framework surrounding health care utilization in general. The ‘Behavioural Model of Health Services Use’ developed by RM Andersen in 1968 (updated in 1995 and renamed the ‘Andersen Health care Utilization Model’) is considered foundational work in this area (Andersen, 1995).

This conceptual model describes the factors that lead to the use of health services. According to the model, there are different dynamics that affect an individual’s health care utilization. Andersen categorizes predictors into three categories: need, enabling and predisposing factors
Briefly, *predisposing* characteristics are those that predispose individuals to use or not use health care services (such as age, sex, ethnicity); *enabling* characteristics are those that either increase or decrease the likelihood of health care service use (such as trust of health care system, income level) and *need* characteristics are those related to health care service, both perceived and actual need, such as the presence of chronic conditions.

The model makes a distinction between equitable and non-equitable access to health care services. Equitable access is driven by predisposing factors and need. Inequitable access is driven by predisposing and enabling factors. For example, an individual who believes Western medical health care services are beneficial to their perceived need are more likely to seek care; however, the ability to access services might vary based on ethnicity, sexual orientation, economic status, and other factors. Need, predisposing and enabling factors, which can change over time, can all affect an individuals’ use of health care services, including regression to the mean.

In general, health care utilization data does not capture the entire population in terms of the outcome of interest as not everyone can or will access services. The studies reported here are aided by the inclusion of a population registry. This registry is a database of all residents in the province with a health care card, regardless of whether services are accessed (~ 98% of population) (eHealth Saskatchewan, 2008). However, even despite the use of a population registry, the theoretical framework of health care utilization must be kept in mind when conducting research studies that make use of administrative health care data.
1.3 Literature review

1.3.1 Health care costing in Canada

Understanding how health care costs are measured is important for health system payers, increasingly so, as gross domestic product expenditures on health care continue to rise. Efforts have been made to standardize methods for health care costing at a national level to facilitate international comparisons (OECD, 2011). Canadian health service organizations report detailed financial and statistical data on their operations using the Standards for Management Information Systems (MIS) developed by the Canadian Institute for Health Information (CIHI). MIS standards include a chart of accounts, accounting principles and procedures, workload measurement systems, indicators, management applications and a glossary of terms (MIS Standards, CIHI, 2015). These standards ensure that different organizations report financial data in a consistent, standardized and comparable way.

Total health care costs can be measured using both direct and indirect costs. Direct costs are those costs borne by the medical system from the perspective of the payer. Indirect costs are a comprehensive measure of the individual, employer and societal costs associated with illness – absenteeism, lost productivity, and staff turnover, among others (Boccuzzi, 2003). Indirect costs to society are difficult to measure and quantify – the current study will focus on direct health care costs.

Even when only direct health care costs are considered, some direct costs are typically not accounted for. For example, administrative costs, health records, information management, capital costs, building maintenance, etc. are not typically accounted for in direct patient care costs. The current studies will consider only direct health care costs directly attributable to
patient care as these costs are readily available in administrative health data; recognizing that the available costs used in the following studies do not capture full, direct health care costs.

In order to calculate direct health care costs there are two primary components to measure: utilization (typically, length of stay and number of visits) and unit costs (Woodchis, Bushmeneva, Nikitovic, & McKillop, 2013). There are two main methods used for determining unit costs – micro or macro-costing methods.

Micro-level costing methods attempt to define costs at the individual/person-level. Micro-costing (bottom-up) requires the researcher to have detailed knowledge about all procedures, care providers and equipment used for an individual patient during an individual health system encounter (Chapko et al., 2009; Woodchis et al., 2013). For example, micro-costing an emergency room visit for acute gastroenteritis would include specifics on nursing time, physician time, sample bottles, linens, laboratory costs, etc. This type of costing data are readily available in health care systems that levy user fees for each health care encounter to the patient. Micro-costing health information in Canada, however, is difficult and labour intensive to obtain given the universal nature of the Canadian health care system with few to none user payment models.

Macro-costing (standard costing or top-down) is an aggregated approach to costing health services. Macro-level costing methods use aggregate costing data and allow analysts to determine how much spending is directed to broad categories of health services. The macro approach requires the researcher to have knowledge about the overall health service expenditures and the total number of patients who received services. For example, if an emergency department had an annual expenditure of $3 million and 3,000 patients were seen, the average cost would be $1,000 per patient. Typically, this method is further refined by grouping patients based on
similar diagnoses, age, sex, and other characteristics to create relative weights – rather than simply dividing total costs by total (likely, not similar) patients (Chapko et al., 2009).

The choice as to the most appropriate costing approach will depend on the particular study question(s) at hand (Kiivet et al., 2013). Given that the objectives of the current studies are to understand high-cost use at the individual-level, as well as factors associated with high-cost use in the provincial population, where available, I chose to use a micro-level, person-level costing approach. In addition, one of the main benefits of individual-level costing methods is the “ability to identify persons whose cost profile may be affecting the average cost” (Guilcher, Bronskill, Guan, & Wodchis, 2016) – a direct objective of the current studies.

In Canada, hospital, prescription drug and physician services account for the largest proportion of total health care costs (> 60%) (CIHI, 2015). Direct cost measurement methods are a function of the particular data set being used; specific information on the costing models for each of the datasets used in this thesis are described below.

1.3.1.1 Hospitalization costs

The Discharge Abstract Database (DAD) is a national data repository of all acute care institution separations (discharges, transfers, deaths, sign-outs) in all provinces and territories of Canada, with the exception of Quebec. DAD records include administrative, demographic and clinical data and are submitted to CIHI. CIHI outlines reporting standards, including coding standards and data quality measures, for provinces and territories to adhere to. CIHI conducts routine data quality audits (involving re-abstraction from hospital charts); data quality is consistently good (Resource Indicators, DAD, CIHI, 2015).
For hospitalization costs, CIHI measures actual hospitalization costs (micro-level/bottom-up) in six participating Canadian hospitals in Alberta, British Columbia and Ontario. The ability to conduct individual-level costing requires complicated software and significant staff time; CIHI takes this individual-level costing data from participating reporting institutions and creates products to estimate individual-level costs for use by other Canadian acute care institutions. All non-participating institutions (including all hospitals in Saskatchewan) use CIHI products – resource intensity weights, case mix grouping, cost per weighted case and cost of a standard hospital stay – to obtain individual-level cost estimates for their institutions. CIHI is actively recruiting additional hospitals to improve the representativeness of the data (CIHI, 2011).

The Canadian Institute for Health Information classifies all hospitalizations in the Discharge Abstract Database into Case-Mix Groups (CMG)™. Case-Mix Groups are grouped based on the most responsible diagnostic code, in addition to resource patterns (procedures or interventions received). In Canada, currently, CMG diagnostic codes use the World Health Organization’s International Statistical Classification of Diseases and Related Health Problems, 10th Revision (ICD-10-CA) (World Health Organization, 2011). Within each CMG, cases are further categorized by age categories.

The case mix methodology takes individual diagnostic codes and groups similar diagnoses. This is useful not only for understanding patient groups by clinical profiles, but, is a key component to calculating Resource Intensity Weights (RIW). The RIW is a weighted measure of the relative cost of a hospitalization whose value is then used with health service expenditure data to calculate hospitalization costs at the individual-level. For example, in Canadian hospitalization data, RIWs are assigned to each CMG; which can then be used to measure the relative costs associated with different diagnostic, surgical procedure and demographic characteristics. Each
CMG has a base RIW and individual variation is accounted for by multiplying the base RIW by the individual patient’s actual length of stay, co-morbidity level and interventions received. For example, a patient with an RIW of 2.0 has used twice as many resources as a patient with an RIW of 1.0. For each of the study subjects in the following chapters, individual-level hospitalization costs per person were calculated by multiplying the CIHI RIW value at the patient-level by the CIHI derived value ‘cost of a standard hospital stay’. Readers will recall this standard value is derived by CIHI using micro-level costing data collected from the six institutions in Canada who report all resources used for each individual-level patient hospital encounter.

1.3.1.2 Emergency department costs

Costs associated with each emergency department (ED) visit in Saskatchewan are a function of whether or not the facility reports to the National Ambulatory Care Reporting System (NACRS). For NACRS reporting facilities, the costing method is very similar to the methods outlined above for hospitalizations; each visit is assigned a CIHI derived RIW, the cost of a standard emergency department visit is calculated by CIHI using the micro-level costing data provided by the six reporting facilities in Canada and the values are multiplied to determine individual-level costs. With ED visits, however, Case Mix Grouping methods are not applied, an RIW is assigned by individual ‘chief complaint’ coded at ED discharge.

In Saskatchewan, however, not all EDs are NACRS reporting facilities. This is mainly due to human resource capacity and available knowledge of ICD coding and electronic infrastructure limitations, largely in rural Saskatchewan. All EDs in Saskatoon, Regina and Prince Albert report to NACRS. North Battleford has one ED reporting to NACRS and one that does not. All
other EDs in the province (n = 6) do not report NACRS. As a consequence, the Saskatchewan Ministry of Health assigns costing values to ED visits at all NACRS non-reporting facilities.

1.3.1.3 Long-term care costs

In Saskatchewan, long-term care (LTC) costs are created using a macro-level costing method. The Community Care Branch, Saskatchewan Ministry of Health assigns a cost to each long-term care resident based on the total long-term care budget divided by the total number of LTC beds. This value is then adjusted, per fiscal year, by the number of days an individual is resident of a LTC facility; this value is then assigned to each LTC resident.

1.3.1.4 Prescription drug and physician billing costs

Prescription drug databases and physician billing databases in Saskatchewan provide the actual dollar figure associated with each service. Where total costs are delineated into costs borne by the provincial government and costs borne by the individual patient; only government costs were included in the calculations in the proceeding studies (excludes all out-of-pocket payments). In Saskatchewan, the majority of physician visits cost $32; this value will fluctuate depending on the service provided. As the federal government is responsible for payment of prescription drug costs for a portion of the Canadian population, prescription drug costs in the following studies are limited by missing all federal beneficiaries (Registered Indian, RCMP prior to 2013 and federal inmates ~19% of the provincial population). This limitation in federal versus provincial coverage applies to costing data in prescription drug data, only.
1.3.2 Definition of 'high-cost' health care use

Methods to determine who is a high-cost user in health care vary. Most studies, after person-level costing methodologies have been employed, simply sum the total costs over the period of interest by unique individual and sort them from highest to lowest – the top 1%, 5% or 10% are deemed ‘high-cost’. A criticism to this approach has been that it oversimplifies a heterogeneous group of high-cost users and does not take into account a population threshold value above which an individual would be considered ‘high-cost’.

Instead, the following studies take the approach recommended by Wodchis et al. Briefly, this methodology recommends identifying unit costs associated with individual health care utilization data from administrative databases and summing the values. This then provides a measure of direct health care costs incurred by government at the individual-level; Wodchis refers to this as a ‘person-level costing methodology’ (Wodchis, Bushmeneva, Nikitovic, & McKillop, 2013; Guilcher et al, 2016).

Following the person-level costing methodology, population threshold values are then calculated. Population threshold values refer to the dollar amount associated with the 90th percentile of health care costs per fiscal year of total health care costs for the entire provincial population, not just those under study. Once this dollar amount is calculated, then, within each study cohort(s), any individual exceeding this dollar amount, per fiscal year, will be considered ‘high-cost’ (Wodchis, Austin, & Henry, 2016).

‘High-cost’ use can also be defined as individuals who exceed the 90th percentile, 95th percentile or 99th percentile of total population costs. I chose to focus on individuals in the top 10% of total costs per fiscal year; the reasons for this are two-fold. First, the majority of high-cost user studies
in the literature (see section 1.3.4) define ‘high-cost use’ as those in the top 10% of costs (individuals exceeding the 90th percentile of costs). Second, one of Canada’s foremost experts in high-cost health care research, Dr. Walter Wodchis, generously donated his time and expertise to a portion of the conceptualization of the studies in this thesis; Dr. Wodchis’ work typically defines ‘high-cost’ use as the top 10% of costs. The arbitrary nature of any cut-off value used, however, is acknowledged.

To further complicate high-cost use definitions, categorizing an individual as being a ‘high-cost user’ at one point in time is less than ideal. As evidence, high-cost user’s costs will decline naturally over time even without intervention, something that makes sense when thinking about a catastrophic illness resulting in high health care needs followed by recovery (resulting in reduced costs). As described above, Chakravarty and Cantor borrowed a statistical term, regression to the mean, to describe this phenomenon among high-cost health care users (Chakravarty & Cantor, 2016). Interestingly, their quantitative review of the problem demonstrated that high utilizers of emergency department services were less likely to experience such regression to the mean when compared to individuals with frequent hospitalizations. This may seem logical – individuals with frequent hospitalizations likely have a catastrophic illness that is more likely to be transient (recovery or death) compared to high users of emergency departments; in addition to hospitalization occurring at the discretion of health care providers with clinical judgement while visits to an emergency department typically occur when an individual chooses to access services.

The larger concern, however, in defining ‘high-cost use’ at one point in time, is that individuals who are persistently high-cost year after year are likely different from individuals who are episodically high-cost. Where data availability permitted (Chapters 2 and 3) I ensured that both persistent and episodic high-cost use was examined separately.
1.3.3 Statistical modelling of health care costs

The majority of individuals do not use health care services in a given year (zero values); this zero inflation should be taken into account when modelling health care costs (Mullahy, 2009; Tooze, Grunwald, & Jones, 2002). In addition, costs are typically positively skewed due to the very few patients who incur very high costs (Basu & Manning, 2009). Previous studies have compared different statistical modeling approaches to health care costing, and in general, recommend an approach that is based on the assumptions in agreement with the distribution of the costing data.

Dodd and colleagues, using an example of costs incurred during the treatment of inflammatory bowel disease (IBD), compared different multivariable regression techniques to analyze health care cost data. The authors compared normal and bootstrapped multiple linear regression, median regression, gamma model with the log link and normal linear regression of log costs; gamma modelling with the log link was found to be the most suitable model (Dodd, Bassi, Bodger, & Williamson, 2006). As the authors focused on only positive costs associated with IBD they did not have zero values in their data. Dunn and colleagues, focusing on mental health costs, provide recommendations on modelling approaches to use and argue modelling techniques should vary depending on whether the aim is to describe, explain or predict future costs. They go on to recommend that if the primary aim is to explain, generalized linear models (with careful attention to the appropriate error distribution given the preponderance of zero values) is best. For prediction, they recommend regression on raw costs using ordinary least-square methods; cautioning that whatever approach is taken should be mindful of the robustness of the method due to incorrect distributional assumptions and to not automatically assume bootstrapping methods will solve this issue (Dunn, Mirandola, Amaddeo, & Tansella, 2003).
A recent study, focused on predicting individuals at risk of frequent emergency department visits and health care costs, employed machine learning techniques to analyze free text data contained within electronic medical health records. The authors conclude their modelling was more predictive of patients who will generate future high costs than future emergency department visits. Further, as is relatively common with predictive modelling, they go on to state that “it remains to be seen if these predictions can be used to reduce costs by early intervention in this cohort of patients” (Frost et al., 2017). Another recent study, focused on simulating various scenarios for health care cost savings specific to the implementation of ‘accountable care organizations’, recommends the use of agent-based modelling. Authors note, “[agent-based modelling] is an increasingly important method to construct a health care analytics model that can help inform health policy and health care management decisions” and note, despite their findings, there is a relative absence of published studies employing the method (Liu & Wu, 2016).

As the studies in this thesis focus on a binary outcome the aforementioned modeling techniques and suggestions are interesting, but, less than relevant. At the outset of this work, using the findings and recommendations above, I employed different ways to categorize the costing outcome – as a continuous measure using linear regression, zero-inflated count models and others. However, through the course of this work, it became clear that the most interpretable categorization of the outcome was binary – high-cost versus not high-cost. I focused, then, on the best practices and guidance for logistic regression modelling as described by Hosmer and Lemeshow (Hosmer & Lemeshow, 2005), and Classification and Regression tree analysis as described by Breiman and colleagues (Breiman, Friedman, Stone, & Olshen, 1984).
1.3.4 High-cost health care literature

Generally accepted to be at the forefront of high-cost health care research is Dr. Jeffery Brenner and his team in Camden, New Jersey (Brenner, 2013; Kaufman et al., 2014). Camden, with a population of 77,334 is one of the poorest cities per capita in the United States. Brenner and his team use the term ‘hot-spotting’ to describe their methods to identify areas of high health care use in Camden. ‘Hot-spotting’ is a term borrowed from computer science – hot-spots are pieces of code that take longer to run or bottleneck processes (Blandy, Afif Saba, & Urquhart, 1999). Brenner and colleagues used the term to define frequent utilization of health care services. The New Jersey researchers linked hospital admission and physician billing data, including costing, to identify hot-spots. For example, one building (n = 615) in the poorest neighbourhood in Camden was found to account for 3,901 visits, 615 patients, $83 million in charges ($21,000/visit), $12 million in receipts and a 15% collection rate for payment (Kaufman et al., 2014).

In September 2018, a systematic review of high-cost health care user studies was published. Study authors searched specific key words and included studies reporting characteristics and utilization within the top health care spenders. The systematic review found the following: high-cost users had a higher prevalence of multiple chronic conditions, higher prevalence of mental illness, health care costs increased across the lifespan but 50% of high-cost users were less than age 65, and, higher income was associated with higher costs in the US, but, with lower income in all other geographies studied. Additionally, preventable spending was estimated to be, at most, 10%, the top 10% accounted for 68% of total costs, persistent high-cost users varied between 24 to 48%, and, no more than 30% of high-cost patients were end-of-life (Wammes et al., 2018).
A recent study compared high-cost users (top 10% of Medicare Fee-for-service beneficiaries) and ‘hot-spotters’ (those in the top 10% of costs and ≥ four hospitalizations or emergency department visits during the study period). The study, entitled “High-cost patients: Hot-spotters don’t explain the half of it” found that hot-spotters represented 9% of the high-cost population, accounted for 19% of their overall costs and 55% were in the ‘frequent care’ grouping; psychiatric disorders were most prevalent in the frequent care group (Lee, Whitman, Vakharia, , & Rothberg, 2017). The authors recommend future research to better define high-cost users, and not simply use definitions reliant on high utilization (‘hot-spotters’). Given health care decision-makers’ interest in cost savings/expenditures, the expertise of my collaborators, and, the recommendation from Lee et al, I chose to focus the studies in this thesis on ‘high-cost’ users, as opposed to frequent use.

1.3.4.1 American high-cost user studies

The majority of high-cost user studies have been conducted in the United States (US). Given that the US health care system is not universal to the population as a whole, the findings are not always applicable in the Canadian context. Despite this, overall findings between American and Canadian studies have been similar.

Quantitative studies on high-cost health care use, in both the US and Canada, have generally found high-cost use to be associated with one or more of the following: 1) complex, multiple chronic conditions (multi-morbidity), 2) catastrophic illness (for example, motor vehicle accident or major cancer), 3) high costs at end-of-life, 4) mental health and addictions, 5) institutional
living, and 6) various indicators of lower socio-economic status (for example, food insecurity and poverty) (see Section 1.3.4 and Wammes et al., 2018).

American high-cost use studies typically analyse data available from a specific health management organization or specific government insurance coverage, such as Veteran’s or Medicare recipients. Pieces of the high-cost use puzzle have been discovered in separate studies using different patient populations and different methodologies. An overview of these findings follows.

Fleishman and Cohen, using American survey data, demonstrated the importance of the number of comorbid chronic conditions associated with high-cost use (Fleishman & Cohen, 2010).

In a Florida Medicaid study, Robst demonstrated the importance of accounting for institutional living (analogous to long-term care residents in the proceeding studies) and prior mental health and addiction health care costs when predicting future persistent high-cost users; however, Robst was limited to mental health clients only (Robst, 2015). Lu and colleagues, in Virginia, found that a count of chronic conditions, demographic data, and, specifically flagging individuals who had experienced congestive heart failure showed the best model performance for understanding variables associated with high-cost use (Lu, 2015). In 2016, Chang and colleagues further developed the understanding around high-cost use by taking into account prescription medication costs. In fact, they found that models containing only drug-specific data performed better than models which accounted for diagnostic groupings, medication and prior health care costs when predicting consistent high-cost health care users (Chang et al., 2016).

InterMountain Healthcare in Utah is a privately run health care centre, often considered on the forefront of implementing research findings into practice in the US. Researchers Wrathall and
Belnap, using InterMountain patient data, built upon previous American findings and argue their logistic model containing clinical characteristics, drug data and socio-demographic characteristics performed best in both explaining and predicting future high-cost use – a model that is actually used in practice in order to provide comprehensive health care interventions to high-cost patients and individuals at risk of future high-cost use (Wrathall & Belnap, 2017).

Further refining Wrathall and Belnap’s models, a recent study from the US found that population segmentation methods were useful for defining actionable high-cost user cohorts (Joynt et al., 2017). The six cohorts the authors defined were: under 65 years of age and disabled/end-stage renal disease; frail elderly; major complex chronic, minor complex chronic; simple chronic; and relatively healthy. Similar to the study I conducted (Chapter 4), individuals in the top 10% of spending were considered high-cost. The authors concluded that using simple criteria to segment, in this case, Medicare beneficiaries, into meaningful subgroups is a useful method to target interventions – a conclusion similar to the work described in this thesis (Chapter 4). Besides only using data specific to one population (Medicare recipients), and, a non-validated approach to segmentation, this study is further limited by the fact that ‘high-cost use’ was defined overall and not within each population segment. In addition, the authors did not conduct analyses beyond descriptive statistics, limiting the ability of their work to point to risk factors associated with high-cost use within the population segments.

Another recent study also based in the US also described the utility of population segmentation in describing high-cost users. This study employed density-based cluster analysis to determine the population segments; their analysis was limited by not including any indicators relating to socio-economic status (Powers et al., 2018).
When defining high-cost use and designing interventions around a relatively small, homogenous population, such as the one in Camden, New Jersey (Brenner, 2013; Kaufman et al., 2014) or InterMountain Health care in Utah (Reiss-Brennan, Briot, Savitz, Cannon, & Staheli, 2010; Wrathall & Belnap, 2017), it is easier for the insurers to create policies aimed at improving health outcomes and inappropriate health care costs. For example, an interdisciplinary team of care providers in Camden were able to co-ordinate their services within a specific geographical location (one apartment building) around a relatively homogenous population (lower socio-economic status). However, when a population is more diffuse and broad, as is the case with publicly insured provincial populations in Canada, describing and creating policies aimed at reducing high-cost health care use becomes more difficult.

1.3.4.2 Canadian high-cost user studies

The first published high-cost health care user studies in Canada were conducted in Manitoba (Roos et al., 2003; Roos et al., 1989). These studies were limited in that they focused on the elderly population; however, they did provide solid evidence to show that, in a cohort of elderly ‘high-cost users’ over a 16-year timeframe, individual persons do not necessarily remain ‘high-cost’ throughout the study period. The authors suggest this demonstrates the need for interventions targeted at the sub-population who would benefit most; I would argue the findings further reinforce the need to understand the underlying risk factors associated with high-cost use.

Recent Canadian high-cost user work has been largely focused in the province of Ontario. A seminal study on high-cost health care use by Wodchis et al found that high-cost users were a heterogeneous group, that approximately 30% of high-cost users in any given year go on to be high-cost users in subsequent years, and, that institutional care was the costliest group of high-
cost users (Wodchis et al., 2016). This descriptive analysis in Ontario built a foundation for subsequent studies on high-cost health care use in Canada.

Tarasuk and colleagues were the first to show the link between household food insecurity and health care spending in Canada. Total health care costs were linked, at the individual level, with Canadian Community Health Survey data in Ontario. Independent of other social determinants of health measured, food insecurity was predictive of future health care costs (Tarasuk et al., 2015).

Guilcher and colleagues (Guilcher et al., 2016), building on the person-level costing methodology developed by Wodchis et al (Wodchis et al., 2016), worked to define person-centered episodes of care based on costing data. The authors linked administrative health databases in Ontario; however, all analyses focused only on individuals at the point of an acute care hospitalization. This paper provides important methodological considerations for high-cost health care use but does not adopt a population health perspective and therefore does not address modifiable factors associated with the outcome.

Rosella and her team in Ontario have been investigating high-cost health care users using a population health approach (Alberga et al, 2018; Fitzpatrick et al., 2015; Rosella et al., 2014). Their first study linked Canadian Community Health Survey data to physician billing data in Ontario. In a population of high-cost users (n = 91,223) authors found that poor self-rated health, lower socio-economic status, former daily smokers, physically inactive, current non-drinkers of alcoholic beverages and obese individuals were more likely to be high-cost users. While one of the first studies to take a population health approach and providing important information on the risk factors associated with high-cost use, this study did not consider any administrative health data beyond physician visits, nor did it provide results by relevant sub-groups, just for the province of Ontario as a whole (Rosella et al., 2014).
Rosella and Fitzpatrick, in another analysis, linked two cycles of the large, representative Canadian Community Health Survey, for years 2003 and 2005, to administrative health data in Ontario and found that measures of low socio-economic status (low income, less than post-secondary education and living in high-dependency neighbourhoods) increased the odds of future high-cost use. After adjustment, they found that the strongest predictors were food insecurity, low personal income and non-homeownership. This study was the first in Canada to link survey data with quantitative administrative health data to investigate high-cost health care use. Their results point to a strong association between high-cost use and the social determinants of health (Fitzpatrick et al., 2015). I would argue, however, that results are less ‘actionable’ to policy and decision-makers due to the inability of the methods used in the study to segment the study population into clinically relevant groups (see Chapter 4).

Rosella and colleagues, most recently, were interested in whether health behaviour risks (as measured by smoking and physical activity) were predictive of future high-costs. The authors linked a combined cohort of two cycles of Canadian Community Health Survey data (2005-2009) to actual health care cost data occurring in the future (following survey response). Compared to individual behaviours, increasing the number of health behaviour risks significantly strengthened the odds of becoming a high-cost user in subsequent years (Alberga et al., 2018).

Thavorn and colleagues in Ontario (Thavorn et al., 2017) found that socio-demographic factors modified the association between high-cost use and multi-morbidity. The authors conclude that both multi-morbidity and socio-demographic factors should be considered in any study aiming to understand high-cost health care use.
In terms of predicting future high-cost use, to the best of my knowledge, there is one published Canadian study (Chechulin, Nazerian, Rais, & Malikov, 2014); an additional Canadian study aimed at predicting a future mental health high-cost user is currently in press (in press: de Oliveria et al). Chechulin et al, developed logistic models to predict persons at risk of future high-cost use in Ontario (excluding individuals < 5 years of age; high-cost use was defined as individuals in the top 5% of costs in a given year). The authors found the strongest predictors in their model were age, chronic health conditions, measures of health care utilization (particularly, mental health inpatient stay), long term care residency and dialysis. In addition, neighbourhood level measurements of material deprivation increased the odds of high-cost use (Chechulin et al., 2014). The authors suggest the model could be used for predicting, but, there is little evidence to suggest it has been used beyond the classical epidemiological sense.

The majority of high-cost user published studies have been conducted in North America. In the United Kingdom, when studied, the population of interest is typically focused on high utilizers as opposed to high-cost use and typically referred to as ‘super-utilizers’ or ‘frequent fliers’. An early study published in the British Medical Journal identified ‘frequent fliers’ in the National Health Service (NHS) as those who had three or more re-admissions to hospital in a one year period. Authors found that the odds of being a ‘frequent flier’ were greater with socioeconomic deprivation, male sex, initial emergency admission and increasing comorbidity (Williams, Bottle, Rogers, & Aylin, 2005).

In Ireland, and not incongruent with the finding of mental health and addiction clients associated with high-cost health care use in North America, a recent study focused on high-cost users of adult intellectual disability residential services in Ireland found that poorer levels of individual psychological well-being were related to higher costs (other factors studied such as staff/client
ratios, number of staff at residential units, and others were, interestingly, not found to be associated) (Cronin & Bourke, 2017).

Swedish researchers and health policy makers have focused their work on a population of ‘complex, high-need, high-cost elderly’. The work has been largely attributed to the story of ‘Esther’; who, after 5.5 hours in the Swedish medical system had provided her medical history 36 times before she was admitted to hospital and provided treatment. The ‘Esther model’ in Sweden, a program designed to co-ordinate clinical care around being mindful about “what is best for Esther” at each decision, is considered one of the pioneering programs of patient involvement in health system improvement (Vackerberg, Levander, & Thor, 2016).

Scottish researchers are generally credited with developing some of the first predictive models, put into practice and demonstrating good results, to assess the risk of hospital re-admission. While not specifically focused on a high-cost high-need population, and results are limited only to those hospitalized, the models are arguably important methodological contributions to the body of literature in predicting future health care needs (Mahmoud, 2016).

1.3.4.3 Qualitative high-cost user studies

To the best of my knowledge, only two qualitative inquiries specifically aimed at high-cost health care use have been published. One recent study interviewed five archetypal high-cost users, their family members and their clinicians in addition to carrying out a quantitative analysis of their claims data and medical charts to explain how patient and health system characteristics influenced health costs (Ganguli et al., 2016). These five patients were heterogeneous, and similar to previous research, the cost drivers for these patients appeared to be complex, active medical issues; physical disability/frailty; mental illness and substance use. The authors describe
six cost drivers which were associated with increased costs and ten cost modifiers that were associated with either higher or lower costs, depending on the individual. These factors were categorized at both the patient-level and health care system level. Cost drivers included: Medical issues, Physical disability/frailty and mental health and addictions (Patient level), and, Care fragmentation, Poor communication, and Unreliable care processes (System level); Cost modifiers included: Socioeconomic status, Insurance coverage, Social network, Patient activation/engagement, and Trust (Patient level), and, Practice style/culture, Medical knowledge, Health resources, Payment structure, and Clinician caring (System level). Of note, the authors conclude that trust was a key theme and that in particular, lack of trust in individual doctors or institutions were cost drivers (Ganguli et al., 2016). While this study provides important information on factors associated with high-cost use that cannot otherwise be ascertained from studies using administrative health data alone; one could argue by interviewing only five individuals their findings are better categorized as elucidating a range of factors that may be of interest, as opposed to their conclusion that there are 16 variables associated with high-cost health care use.

An earlier qualitative study interviewed 19 participants of the Camden Coalition of Health care Providers, in Camden, New Jersey. Here, the authors identified three themes that arose from patient interviews. The first theme centered around adverse childhood experiences (ACE); nearly all those interviewed indicated one or more ACE. The second theme focused on previous perceived negative interactions with the health care system. And the last centered on the positive interactions – the feeling that someone genuinely cared – in the new model of care they were receiving (Mautner et al., 2013).
A multitude of mixed methods and qualitative literature does exist, however, focused on understanding frail elderly, socially vulnerable and multi-morbid populations; populations that arguably are high-cost/high-utilization of health care. Interestingly, these studies typically find good communication and trusting relationships are factors in the care quality experience of people with complex care needs.

### 1.3.4.4 The value of high-cost user research

An opinion piece in the *Annals of Family Medicine* cautions against focusing on high-cost, high-need populations when working to achieve health system transformation (Newton & Lefebvre, 2015). The authors feel that program interventions narrowly focused on high-cost, high-need populations drive system investment and resources away from the hard work of primary care transformation by adding yet another silo and programmatic layer to an already fragmented care system. The authors conclude that the key to intervening effectively on a super-user population will to be to have effective interventions embedded in the family physician office: team-based care, better scheduling services, IT systems to drive improvement and care management, capacity to intervene outside of an office setting and community-based infrastructure to facilitate collaboration between primary care and social services.

However, in the same issue of this journal, the opposite point is argued. The author here states that by having a concerted strategy aimed at the ‘super-utilizer’ population one can unlock the key to health system transformation. The intersection of both poor quality health care, inadequate health outcomes and high health care costs, she feels, can genuinely engage health system decision-makers towards real system change (Emeche, 2015).
Given intervention research into high-cost populations in Camden, New Jersey and InterMountain Healthcare in Utah have demonstrated success by decreasing costs and increasing health/quality of life, I am of the opinion that, while not likely to be the ‘answer’ to health system transformation, understanding this universally small group of patients consuming the majority of health care resources is one piece of the puzzle for improving a complex system.

1.3.4.5 Summary

From 1985 (n = 79) to 2018 (n = 1,198) the number of PubMed citations involving the term ‘high-cost use’ has increased over 15 times (PubMed - NCBI, 2018). Given rising health care costs, increased demand from larger elderly populations with multi-morbidity, and, health consuming more and more of the total gross domestic product in many countries around the world, policy makers and health researchers have, reasonably, been looking to the population of ‘high-cost users’ for cost savings.

1.3.4.6 How my thesis advances high-cost user research

The following studies in this thesis build upon the aforementioned best practices in individual-level health care cost attribution, methodological considerations in health care costing data, definition of ‘high-cost’ use, and, the body of literature surrounding risk factors associated with high-cost health care use. One of the strengths of the studies that follow include the benefit of building upon the body of literature describing risk factors associated with high-cost use and ensuring they were accounted for, as best I could, in administrative health data. Additionally, literature typically accounted for only a few risk factors whereas I made every effort possible to include as many as possible. For example, specific risk factors, such as, long-term care
residency, multiple chronic conditions, prescription drug costs, socio-economic variables and mental health and addictions clients were identified in my study cohorts.

Findings from qualitative studies, however, were more difficult to account for in the following quantitative studies. For example, I was unable to determine, using routinely collected administrative health data, the level of trust an individual had in their primary care provider. As the best available proxy for this finding, I used an indicator in Canadian physician billing data that indicates the degree of ‘connectedness’ an individual has to their primary care provider (‘usual provider continuity index’; see Chapter 2.3).

In addition to accounting for the previous data found to be associated with high-cost use to the best of my ability, I also, to the best of my knowledge, created novel variables when exploring high-cost use. For example, the studies conducted in Chapter 2 and 3 are the first to include an indicator of ‘unstable housing’ at the individual level, which I defined as frequent address changes using the provincial shared client index and street outreach clients; databases that, to the best of my knowledge, had never before been used for research in Saskatchewan. The study in Chapter 4 is, to the best of my knowledge, the first in Canada to make use of the newly developed ‘made in Canada’ approach to population segmentation.

The Canadian Institute for Health Information (CIHI) recently (April 2018) released SAS code to all provinces/territories designed to group all provincial residents, based on their past 24 months of health care utilization, into clinically meaningful sub-populations. The resulting ‘Population Grouping Methodology’ was the result of years of technical collaborative scientific working groups, which included leading Canadian clinicians. Extensive validation exercises have been conducted using Alberta, British Columbia and Ontario data (unpublished; available from CIHI
upon request) but to the best of my knowledge the method has not yet been used to further understanding of high-cost health care use, and, had never before been used in Saskatchewan.

Given the importance of the Population Grouping methodology in grouping the provincial population into clinically meaningful sub-groups in Chapter 4, and, the relative ‘newness’ of the method and lack of information in the published literature, I provide the following details on CIHI’s Population Grouping Methodology.

1.4 Organization of thesis

This thesis is comprised of an introductory chapter, three quantitative studies which address the key objectives, and, a concluding chapter. The reader will note that the population under study changes across the three manuscripts. A short explanation for this is warranted.

This research was conducted in near real-time, and, in the case of the majority of the study objectives, in direct response to requests for research findings from provincial-level decision-makers. First, prior to my involvement in the work, at the request of the provincial Ministries of Health, analyses conducted in both Saskatchewan and Ontario demonstrated that mental health and addiction (MHA) clients were among one of the highest cost users of the health care system in both provinces. As a direct result, a Canadian Institute for Health Research (CIHR) team grant was awarded to these researchers to understand mental health and addiction high-cost use in the two provinces.

When the CIHR ‘Quick Strike’ team grant was awarded the team added me to the grant as a student researcher and permitted my PhD thesis to form a part of the overall team grant. This
initial grant-based project was focused solely on mental health and addiction clients in Saskatoon, Saskatchewan and in Ontario – as mental health was known to high-cost and pressure was mounting to provide evidence to inform provincial decisions for newly announced federal funding in mental health and addictions. At the point I was brought on to the project I led and conducted the work to ensure that in Saskatchewan: 1) disparate databases that would be critical to understanding high-cost mental health and addiction clients were included, and, 2) descriptive analyses geared towards decision-maker requests were completed.

As the research progressed, however, it became clear that a broader understanding of high-cost use, from the perspective of a provincial health care insurer, was needed in order to adequately address decision-makers’ questions – mental health and addictions was a piece of the puzzle but not the only one. To that end, the population under study in my last analytical Chapter focuses not only on mental health and addiction clients in Saskatoon, but, on the population of Saskatchewan as a whole.

In order to effectively analyse provincial-level data, I became increasingly interested in the ability to segment the provincial population into clinically meaningful sub-groups.

Simultaneously to my use of the newly developed Population Grouping methodology (‘Pop Grouper’) to address this thinking, two American studies (see section 1.3.4) were published; similarly demonstrating that population segmentation is critical to appropriately understanding high-cost user populations. I partnered with scientists at CIHI who developed the Pop Grouper to conduct the research described in Chapter 4. This partnership included co-delivering a workshop on the Population Grouping methodology for analysts, epidemiologists, and researchers in Saskatchewan.
As the reader will find, however, this progression in populations under study was a natural extension of the findings; the seemingly disparate manuscripts are, interestingly, related and provide reasonable evidence and methods to understand factors associated with high-cost health care use in Saskatchewan (which I would argue may generalize to other Canadian provincial populations). My first study, Chapter 2, taking into account other potentially confounding factors, was the first study in Canada to demonstrate, at an individual-level (as opposed to ecological-level) the importance of unstable housing in high-cost use among mental health and addiction clients; in addition to demonstrating the importance of understanding episodic versus persistent high-cost use.

My second study contributed novel findings to the relatively sparse literature on predicting who is at risk of high-cost health care use in the future. In this study, I deliberately focused on data readily available in Canadian administrative health databases, such as health care utilization variables, to create variables of potential predictive importance; thus demonstrating the relative ease by which well calibrated and discriminatory predictive models can be created at the provincial level using routinely collected data. In addition, by employing a Classification and Regression Tree methodology – a visual output readily understandable to decision-makers – I aimed to make the results accessible, and, of use in practice.

Despite having the same cohort under study in Chapters 2 and 3 (mental health and addiction clients in Saskatoon, Saskatchewan), I employed different methodologies (logistic versus classification and regression tree modelling) for the differing study objectives (explanatory versus predictive modelling), and, different study exclusion criteria.

My last study builds upon my previous findings and, I feel, addresses a gap in the high-cost user literature by contributing a useful method to segment provincial populations to understand high-
cost users in general. The CIHI Population Grouping methodology is showing promise in segmenting Canadian provincial populations by disease categories – Chapter 4 provides additional details on this method. This newly developed method also includes hospital costing estimates; however, instead of employing the Pop Grouper costing method, for consistency across the three studies, I chose to use the same traditional micro-level, person-level costing method across the three quantitative studies. The results of this ‘On defining actionable high-cost health care use’ study are particularly compelling and circle the reader back the important findings in Chapter 2 – socio-demographic variables were not statistically significantly associated with high-cost health care use overall, however, once population segmentation methods are employed (addressing the natural heterogeneity of the high-cost user population) socio-demographic variables (in this case, low neighbourhood income) were found to be statistically significantly associated with high-cost health care use in some health profile groups.

The quantitative studies in this thesis (Chapters 2, 3 and 4) have been written as stand-alone manuscripts. Each manuscript is currently either published, or, under peer review. Given this manuscript format, the reader will notice some natural repetition across the various stand-alone chapters (for example, summaries of previous high-cost health care user literature, description of administrative health databases and person-level costing methodology). Additionally, journal word count restrictions limited the ability to provide detailed methodological notes that should appropriately be included in a thesis. Detailed methodological notes appear in ‘supplemental material’ following each Chapter. These details are provided to a) permit an independent analyst with appropriate access to all relevant data to recreate all analyses conducted and b) provide transparency to the methods chosen. All SAS code, including my comments, is available to any interested reader.
Ethical review and approval for this research was obtained from the University of Prince Edward Island Research Ethics Board, the University of Saskatchewan Biomedical Research Ethics Board and the Health Canada and Public Health Agency of Canada Ethics Board for Research Involving Humans.
1.5 References


https://doi.org/10.1016/j.amepre.2015.02.018

https://doi.org/10.1111/j.1475-6773.2009.01080.x

https://doi.org/10.1016/j.amjmed.2016.12.008


administrative data. *Health Policy (Amsterdam, Netherlands)*, 112(1–2), 110–121.
https://doi.org/10.1016/j.healthpol.2013.05.011

https://doi.org/10.2105/AJPH.2017.303655

https://doi.org/10.1177/1355819615577711

https://doi.org/10.1007/s11606-016-3790-3

https://doi.org/10.1007/s10729-014-9279-x


Health System Performance Research Network website:


https://doi.org/10.2196/jmir.8508
2. EFFECT OF SOCIO-DEMOGRAPHIC FACTORS ON HIGH-COST MENTAL HEALTH AND ADDICTION HIGH-COST USE: A RETROSPECTIVE, POPULATION-BASED STUDY IN SASKATCHEWAN


2.1 Abstract

Objective
A small proportion of the population accounts for the majority of health care costs. Mental health and addiction (MHA) patients are consistently high-cost. I aimed to delineate factors amenable to public health action that may reduce high-cost use among a cohort of MHA clients in Saskatoon, Saskatchewan.

Methods
I conducted a population-based retrospective cohort study. Administrative health data from fiscal years (FY) 2009–2015, linked at the individual level, were analyzed (n = 129,932). The outcome of interest was ≥ 90th percentile of costs for each year under study (‘persistent high-cost use’). Descriptive analyses were followed by logistic regression modelling; the latter excluded long-term care residents.

Results
The average health care cost among study cohort members in FY 2009 was ~ $2,300; for high-cost users it was ~ $19,000. Individuals with unstable housing and hospitalization(s) had increased risk of persistent high-cost use; both of these effects were more pronounced as comorbidities increased. Patients with schizophrenia, particularly those under 50 years old, had
increased probability of persistent high-cost use. The probability of persistent high-cost use decreased with good connection to a primary care provider; this effect was more pronounced as the number of mental health conditions increased.

**Conclusion**

Despite constituting only 5% of the study cohort, persistent high-cost MHA clients (n = 6,455) accounted for ~35% of total costs. Efforts to reduce high-cost use should focus on reduction of multi-morbidity, connection to a primary care provider (particularly for those with more than one MHA), young patients with schizophrenia, and adequately addressing housing stability.

### 2.2 Introduction

Evidence has long demonstrated that a small proportion of the population (<10%) accounts for the majority (50–70%) of total health care spending (Densen et al. 1959): individuals commonly referred to as ‘high-cost users’. As early as 1988, Taube et al. demonstrated that individuals with mental disorders comprised 9% of high-cost users in 1 year (i.e., those with ≥25 visits), but accounted for nearly half (49.2%) of total outpatient expense (Taube et al. 1988). In a Medicaid population study of high-cost users, Buck et al. (2003) showed that people using mental health and substance abuse services constituted 11% of all enrollees but accounted for approximately 30% of all high-cost users (Buck et al. 2003). Hunter et al. (2015) found that nearly half of high-cost users had a mental health condition (Hunter et al. 2015).

Three Canadian studies have focused on patients with high costs and mental health and addiction (MHA) issues. de Oliveira et al. demonstrated that high-cost MHA patients incur 30% more health care costs per capita compared to high-cost users with no mental health conditions (de Oliveira et al. 2016b); a subsequent study demonstrated that MHA high-cost patients (‘MHA
high-cost’ patients defined as individuals for whom MHA services accounted for ≥ 50% of their total health care costs) had health care costs 40% higher than those with no MHA-related costs (de Oliveira et al. 2017a). Using a combination of mood, substance use, psychotic and anxiety disorders as the definition of mental illness, Hensel et al. found that rates of mental illness were 39.3% in the top 1% costliest users (compared to 21.3% in the lowest cost group) (Hensel et al. 2016).

The disease burden of mental health and addiction disorders is large; it is estimated that in any given year, one in five Canadians will experience a mental illness or addiction, and, by age 40, one in two will have, or have had, a mental illness (Lim et al. 2008; Ratnasingham et al. 2013; Smetanin et al. 2011).

Mental health and addiction patients can have complex care needs related to their illness, as well as deficits in their social determinants of health, such as housing, income, education and employment (Walker and Druss 2016). Among MHA high-cost users, persons diagnosed with schizophrenia are more often high-cost, largely driven by hospitalizations (de Oliveira et al. 2016a; Fortney et al. 2009; Hensel et al. 2016; Junghan and Brenner 2006; Robst 2012). Stable housing, particularly, has been demonstrated to not only decrease health care service use, which is directly related to cost, but also improve health outcomes overall for mental health and addiction patients (Charkhchi et al. 2018; Fitzpatrick-Lewis et al. 2011; Hwang et al. 2005; Kerman et al. 2018; Rog et al. 2014; Stergiopoulos et al. 2018). Recent studies demonstrated residential instability as a risk factor for high-cost health care use at an ecological level (Thavorn et al. 2017) in addition to psychiatric inpatient cost savings (Rudoler et al. 2018). Given these previous findings, the current study included an administrative database definition of unstable housing in order to understand this covariate at an individual level in the study population.
Using population-based administrative health databases, I conducted an exploratory study to delineate factors amenable to public health action to prevent persistent high-cost use among a cohort of mental health and addiction clients in Saskatoon, Saskatchewan, Canada. I included health care utilization/access measures, such as connection to a primary care provider, and social determinants of health measures, such as unstable housing. The large cohort under study makes use of individual-level data, as opposed to previous studies (Thavorn et al. 2017) which rely on neighbourhood/ecological predictors.

2.3 Methods

Saskatchewan (population ~ 1.2 million) is a Canadian province with a central provincial health insurer. Health data are captured for the provincial population with the exception of residents covered under the federal government (~1%), specifically inmates of federal prisons, members of the Royal Canadian Mounted Police (prior to 2013), Indigenous persons receiving primary health care services on federal reserve, and Canadian Armed Forces (Downey et al. 2006). The study population, due to database availability, was limited to Saskatoon Health Region, the largest health region in the province (n = 360,000).

Databases

Detailed descriptions of Saskatchewan administrative health databases are available elsewhere (Downey et al. 2006). All data were linked at the individual level using the same unique encrypted identifier. Demographic characteristics, location of residence and insurance coverage were extracted from the Personal Health Registration System (PHRS). Hospital data for the province of Saskatchewan, and submitted to the Canadian Institute for Health Information (CIHI), were extracted from the Discharge Abstract Database (DAD) and include all
hospitalizations (including psychiatric hospitalizations) in the province. The International Classification of Diseases (ICD), 10th revision, Canadian Version (ICD-10-CA) was used in the DAD to record up to 25 diagnoses at admission, including the most responsible one. Data on physician services are contained in the Medical Services Database. Physicians paid on a fee-for-service basis submit billing claims to the provincial health ministry; a single diagnosis using a three-digit ICD-9 code is recorded on each claim. Salaried physicians are required to submit ‘dummy’ claims for administrative purposes (shadow billing) but compliance is low; therefore, a level of under-reporting for salaried physician claims will exist. According to a recent CIHI report, a ‘relatively small percentage of Saskatchewan physicians are compensated through salaried arrangements’ (Canadian Institute for Health Information 2008). Residents of long-term care facilities were defined according to the provincial Resident Assessment Instrument-Minimum Dataset (RAI-MDS) for long-term care facilities (RAI-LTC 2.0). The following datasets were available for Saskatoon Health Region only (n = 162,566 individuals): emergency department data recorded in National Ambulatory Care Reporting System (NACRS); the Shared Client Index (SCI), an identity management system used to manage patient contact information; and the Population and Public Health, Saskatoon Health Region, Street Outreach Database. A combination of frequent address changes in the Shared Client Index and individuals in the Street Outreach Database permitted an administrative database definition of ‘unstable housing’. Total government health care costs by publicly funded source (hospital, physician, prescription drug and long-term care) were provided by the Saskatchewan Ministry of Health. Data were linked at the individual level using a unique non-identifiable health services number generated by eHealth Saskatchewan.
Study design

I conducted a retrospective population-based study using administrative health data. The study population included any individual with continuous provincial insurance coverage, resident of Saskatoon Health Region, ≥ 18 years old, alive as of April 1, 2009 (study baseline) and with at least one mental health or addiction-related International Classification of Disease Codes (ICD-9 or ICD-10) diagnosis in any database from April 1, 2003, to March 31, 2015, with the exception of dementia and dementia-related codes. Total health care costs were calculated by unique individual for each fiscal year 2009–2015.

Exposure

Using a combination of emergency department, physician and hospital data, mental health and addiction ICD-10 codes were grouped into the following: substance-related disorders; schizophrenia, delusional and non-organic psychotic disorders; mood/affective disorders; anxiety disorders; and selected disorders of adult personality and behaviour (Table 2.7.6).

Outcome

The primary end-point was persistent high-cost health care use (≥ 90th percentile for each fiscal year 2009 to 2015). Using the costing methodology developed by Wodchis et al., all costs were estimated and assigned to unique study individuals (Wodchis et al. 2013). Briefly, this methodology provides guidance on how to identify unit costs associated with individual health care utilization and how to combine these with utilization data from administrative databases, providing a measure of direct health care costs incurred by government. To account for death during the study period, each study subject’s total health care costs were divided by total number of days observed in the study.
High-cost status among cohort members was determined for each year of the study period using population proportional thresholds (Wodchis et al. 2016). As cost was shown to vary by individual over time, variables of ‘never high-cost’ (never ≥ 90th percentile for any year of the study period), ‘sometimes high cost’ (≥ 90th percentile in any one year) and ‘persistent high-cost’ (≥ 90th percentile for all years, including up to the point of death) were calculated. Total costs were a sum of mutually exclusive government costs for hospitalizations, prescription drugs, emergency department visits, physician services, and long-term care. Physician billing and prescription drug costs were direct totals payable by the provincial health insurer. Hospitalization costs were obtained using the Standard Cost of a Hospital Stay multiplied by the specific resource intensity weight (RIW) at the individual-level (Canadian Institute for Health Information (CIHI) 2015a, b). Long-term care (LTC) costs were defined using the actual government LTC expenditure provided by the Community Care Branch, Saskatchewan Ministry of Health, and calculated as a daily cost per bed. This cost was multiplied by the number of days an individual was resident in LTC. Study cohort costs were further defined as an annual weighted cost, weighted based on the number of days the individual was alive per study year. All yearly costs were adjusted for inflation to FY 2015.

**Covariates**

Variables included were as follows: age, sex, urban/rural location of residence, neighbourhood-level income quintile, death, number of mental health conditions, comorbid conditions, Usual Provider Continuity (UPC) index, unstable housing, and health care utilization measures. All covariates were defined at the point of occurrence (such as death), otherwise, as of study baseline April 1, 2009, permitting a temporal association between exposures (covariates) and persistent high-cost use (outcome).
The UPC index developed by CIHI was used to indicate the degree to which an individual is connected to a primary care provider. Each physician in the province is assigned a unique number. The UPC is a proportion—for each study cohort member, the total number of physician visits to the same general practitioners’ number is divided by the total number of physician visits to all general practitioners over the same time period. For regression modelling, a dichotomous variable of the UPC index score was created using CIHI’s categorization of 0 to < 0.75 (poorly connected) versus 0.75+ (well connected) (Canadian Institute for Health Information 2016).

Neighbourhood-level income quintile was based on dissemination area average household income values from public-use Statistics Canada census files using the postal code methodology developed by the Institut National de Santé Publique du Québec (INSPQ 2007). As hospital stays are known to drive health care costs, ‘hospitalization’ was defined as ever having an acute in-patient stay prior to the costing observation period. Unstable housing was defined as street involvement (street outreach client, n = 967), or ≥ 4 address changes in a 12-month period from 2003 to 2009 (n = 299).

All hospitalizations, emergency department and physician visits (specialty and general practice) from April 1, 2003, to March 31, 2009, were summed by mental health and addiction diagnostic category. As a large proportion of the study cohort had more than one mental health condition (37%), ‘primary mental health condition’ was assigned as the most frequent MHA diagnostic code(s) across physician, hospitalization and emergency department in the study period.

Comorbidities were defined using validated case definitions for chronic disease conditions from the Canadian Chronic Disease Surveillance System (CCDSS): (1) chronic obstructive pulmonary disease (COPD), (2) congestive heart failure (CHF), (3) asthma, (4) diabetes, and (5) coronary
artery disease (CAD) (Public Health Agency of Canada 2003; Feely et al. 2017). See Chapter 2 Supplemental section 2.7.7 for details. For regression modelling, the total number of comorbid conditions was used (0 to 5). Death data were obtained using a ‘verified death file’ (combination of death data from various administrative health databases) created by the Saskatchewan Health Quality Council for research purposes.

Statistical analyses

Following univariate and bivariate analyses, multivariate regression modelling was used to delineate factors associated with persistent high-cost use. Classification and regression tree (CART) modelling facilitated a visual understanding of variables contributing most to persistent high-cost use; subsequent logistic regression modelling, probit link, quantified the relationship between persistent high-cost use and significant co-variates. In logistic modelling, missing values (income quintile and location of residence) were coded as ‘missing’. Sensitivity analyses with and without missing data was conducted.

All long-term care residents (n = 5,435; 4.2% of cohort) were excluded from the logistic modelling. Approximately 36% of persistent high-cost users were residents of a LTC facility. As I aimed to delineate factors amenable to public health action to prevent persistent high-cost use, it was considered reasonable to exclude these study subjects. It was assumed that individuals in LTC facilities met the standardized criteria for requiring their higher level of care and their reason for persistent high-cost use was known.

All analyses were conducted using SAS© Enterprise Guide version 7.1 (SAS Institute Inc. 2017).
The study proposal underwent ethical review and approval by the University of Saskatchewan Biomedical Research Ethics Board, the University of Prince Edward Island Research Ethics Board and the Health Canada/Public Health Agency of Canada Research Ethics Board for research involving humans.

2.4 Results

A total of 129,932 unique individuals eligible for provincial health insurance, resident of Saskatoon Health Region (SHR) and alive as of the study baseline date (April 1, 2009) were identified as belonging to the study cohort (Figure 2.1). Of these, the majority (55%) were ‘never high-cost’; persistent high-cost users accounted for 5% of the study cohort. Cohort characteristics at study baseline are detailed in Table 2.1.
Figure 2.1: Study cohort inclusion/exclusion criteria

Population of Saskatchewan, Canada

\[ n = \sim 1.2 \text{ million} \]

Eligible study population: unique persons, \( \geq 18 \) years, with at least one mental health and addictions ICD-code, excluding dementia, in administrative health databases (physician billing, hospitalization, and/or emergency department) April 1\textsuperscript{st}, 2003 to March 31\textsuperscript{st}, 2015

\[ n = 451,674 \]

Exclude residents outside of Saskatoon Health Region (study population) as of April 1\textsuperscript{st} 2009 (study baseline)

\[ n = 162,566 \]

Exclude those who died prior to April 1\textsuperscript{st} 2009

\[ n = 138,257 \]

Exclude those ineligible for provincial health insurance and/or with greater than a 3 day gap in continuous health coverage

\[ n = 129,932 \]

Exclude long-term care residents (regression modelling only)

\[ n = 124,497 \]
Table 2.1: Mental health and addiction study cohort, Saskatoon Health Region, Saskatchewan (n = 129,932)

<table>
<thead>
<tr>
<th>Covariates (n) (%)/(mean) (SD)</th>
<th>Never high-cost n = 71,834</th>
<th>Sometimes high-cost n = 51,643</th>
<th>Persistent high-cost n = 6,455</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean age, years (SD)</td>
<td>42 (15)</td>
<td>51 (19)</td>
<td>68 (17)</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female (57%)</td>
<td>37,599 (52%)</td>
<td>33,020 (64%)</td>
<td>3,639 (56%)</td>
</tr>
<tr>
<td>Male (43%)</td>
<td>34,235 (48%)</td>
<td>18,623 (36%)</td>
<td>2,816 (44%)</td>
</tr>
<tr>
<td>Location of residence</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Urban (84%)</td>
<td>60,668 (85%)</td>
<td>42,799 (83%)</td>
<td>5,363 (83%)</td>
</tr>
<tr>
<td>Rural (15%)</td>
<td>10,285 (14%)</td>
<td>8,281 (16%)</td>
<td>1,050 (16%)</td>
</tr>
<tr>
<td>Missing* (1%)</td>
<td>881 (1%)</td>
<td>563 (1%)</td>
<td>42 (1%)</td>
</tr>
<tr>
<td>Primary mental health diagnosis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Substance-related disorders (15%)</td>
<td>11,458 (16%)</td>
<td>7,067 (14%)</td>
<td>676 (10%)</td>
</tr>
<tr>
<td>Schizophrenic disorders (5%)</td>
<td>1,025 (1%)</td>
<td>3,126 (6%)</td>
<td>2,235 (35%)</td>
</tr>
<tr>
<td>Mood/affective disorders (27%)</td>
<td>19,594 (27%)</td>
<td>25,928 (50%)</td>
<td>1,810 (28%)</td>
</tr>
<tr>
<td>Anxiety disorders (52%)</td>
<td>39,524 (55%)</td>
<td>1,464 (1%)</td>
<td>217 (1%)</td>
</tr>
<tr>
<td>Disorders of adult personality (1%)</td>
<td>233 (1%)</td>
<td>146 (1%)</td>
<td>27 (1%)</td>
</tr>
<tr>
<td>Number of MH conditions (mean; SD)</td>
<td>1.4 (0.6)</td>
<td>1.6 (0.8)</td>
<td>2.0 (1.0)</td>
</tr>
<tr>
<td>Physician visits per year (mean; SD)</td>
<td>5.5 (5.4)</td>
<td>10.2 (8.6)</td>
<td>24.7 (18.2)</td>
</tr>
<tr>
<td>ED visits per year (mean; SD)</td>
<td>0.2 (0.4)</td>
<td>0.4 (0.9)</td>
<td>1.1 (2.1)</td>
</tr>
<tr>
<td>Number of hospitalizations per year (mean; SD)</td>
<td>0.1 (0.2)</td>
<td>0.3 (0.3)</td>
<td>0.7 (0.9)</td>
</tr>
<tr>
<td>At least one psychiatrist visit (16%)</td>
<td>9,222 (13%)</td>
<td>8,806 (17%)</td>
<td>2,294 (36%)</td>
</tr>
<tr>
<td>Mean health care cost per year ($; SD)</td>
<td>5557 ($745)</td>
<td>3,242($7,467)</td>
<td>28,320($35,003)</td>
</tr>
<tr>
<td>Income quintile</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 (least affluent) (20%)</td>
<td>13,799 (19%)</td>
<td>11,180 (22%)</td>
<td>1,587 (25%)</td>
</tr>
<tr>
<td>2 (18%)</td>
<td>12,633 (18%)</td>
<td>9,033 (18%)</td>
<td>1,165 (18%)</td>
</tr>
<tr>
<td>3 (20%)</td>
<td>14,456 (20%)</td>
<td>10,260 (20%)</td>
<td>1,263 (20%)</td>
</tr>
<tr>
<td>4 (18%)</td>
<td>12,927 (18%)</td>
<td>9,012 (17%)</td>
<td>1,013 (16%)</td>
</tr>
<tr>
<td>5 (most affluent) (18%)</td>
<td>13,113 (18%)</td>
<td>8,833 (17%)</td>
<td>1,028 (16%)</td>
</tr>
<tr>
<td>Missing* (6%)</td>
<td>4,906 (7%)</td>
<td>3,325 (6%)</td>
<td>399 (5%)</td>
</tr>
<tr>
<td>Unstable housing (1%)</td>
<td>412 (1%)</td>
<td>691 (1%)</td>
<td>163 (3%)</td>
</tr>
<tr>
<td>Connection to primary care doctor</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Well connected (43%)</td>
<td>27196 (38%)</td>
<td>25202 (49%)</td>
<td>3056 (47%)</td>
</tr>
<tr>
<td>Not well connected (57%)</td>
<td>44638 (62%)</td>
<td>26441 (51%)</td>
<td>3399 (53%)</td>
</tr>
<tr>
<td>Select comorbid conditions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diabetes (12%)</td>
<td>4,267 (6%)</td>
<td>8,635 (17%)</td>
<td>2,562 (40%)</td>
</tr>
<tr>
<td>Chronic obstructive pulmonary disease (9%)</td>
<td>3,067 (4%)</td>
<td>6,662 (13%)</td>
<td>2,147 (33%)</td>
</tr>
<tr>
<td>Congestive heart failure (7%)</td>
<td>905 (1%)</td>
<td>5,484 (11%)</td>
<td>2,645 (41%)</td>
</tr>
<tr>
<td>Coronary artery disease (11%)</td>
<td>2,346 (3%)</td>
<td>9,535 (19%)</td>
<td>2,843 (44%)</td>
</tr>
<tr>
<td>Asthma (9%)</td>
<td>4,990 (7%)</td>
<td>5,508 (11%)</td>
<td>1,097 (17%)</td>
</tr>
<tr>
<td>Died during study period (FY2009-2015) (7%)</td>
<td>686 (1.0%)</td>
<td>4,249 (8%)</td>
<td>3,515 (55%)</td>
</tr>
<tr>
<td>Long-term care resident</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes** (4%)</td>
<td>13 (0.02%)</td>
<td>3,074 (6%)</td>
<td>2,348 (36%)</td>
</tr>
<tr>
<td>No (96%)</td>
<td>71,821 (99.8%)</td>
<td>48,569 (94%)</td>
<td>4,107 (64%)</td>
</tr>
</tbody>
</table>

*Modelled categorically as ‘missing’ in regression modelling    **Excluded from regression modelling
Compared to never high-cost, persistent high-cost users had higher proportions of females, schizophrenia diagnosis, unstable housing, and deaths. Approximately 30% of high-cost users in FY 2009 went on to be high-cost in subsequent years (data not shown). Persistent high-cost users were less likely to have a primary mental health diagnosis of anxiety and more likely to have a diagnosis of schizophrenia, delusional and non-organic psychotic disorders (Table 2.1).

Individuals with a primary mental health condition of schizophrenia had total health care costs ~ five times higher, and nearly twice the average number of hospitalizations and physician visits, compared to all other MHA diagnoses (Table 2.2).

<table>
<thead>
<tr>
<th>Primary mental health condition</th>
<th>n</th>
<th>Annual health care cost Mean (SD); Median</th>
<th>Annual physician visits Mean (SD)</th>
<th>Annual emergency department visits Mean (SD)</th>
<th>Annual hospitalizations Mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Substance-related disorders</td>
<td>19,201</td>
<td>$1,905 ($5,253); $502</td>
<td>7 (9)</td>
<td>0.4 (1.4)</td>
<td>0.2 (0.3)</td>
</tr>
<tr>
<td>Schizophrenia</td>
<td>5,858</td>
<td>$12,488 ($16,080); $5138</td>
<td>15 (13)</td>
<td>0.5 (1.1)</td>
<td>0.4 (0.7)</td>
</tr>
<tr>
<td>Mood/affective disorders</td>
<td>36,780</td>
<td>$2,251 ($5,346); $847</td>
<td>9 (9)</td>
<td>0.3 (0.7)</td>
<td>0.2 (0.3)</td>
</tr>
<tr>
<td>Anxiety disorders</td>
<td>67,687</td>
<td>$1,765 ($4,195); $647</td>
<td>8 (8)</td>
<td>0.3 (0.6)</td>
<td>0.2 (0.3)</td>
</tr>
<tr>
<td>Disorders of adult personality</td>
<td>406</td>
<td>$2,446 ($6,078); $646</td>
<td>9 (12)</td>
<td>0.4 (1.3)</td>
<td>0.2 (0.4)</td>
</tr>
</tbody>
</table>

Despite comprising only 5% of the population, persistent high-cost users (n = 6,455) accounted for nearly 35% of total costs (FY 2009). Never high-cost users (n = 71,834) accounted for over...
half of the study population but less than 10\% of costs; sometimes high-cost users \((n = 51,643)\) accounted for approximately 40\% of the study population and 57\% of the costs (Fig. 2.2).

Multivariable logistic regression identified predictors statistically significantly associated with persistent high-cost use. Compared to the logit link, a probit link provided a better fitting model, common in studies using health care cost as an outcome and susceptible to heteroscedasticity (Basu et al. 2006). Relevant two-way interactions and one three-way interaction term were retained in the final model as they improved model performance. Age, sex, and primary mental health condition interacted \((p = 0.005)\); patients with schizophrenia, particularly those under 50 years old, had increased probability of persistent high-cost use; females with anxiety had higher
probability of persistent high-cost use compared to males. Neighbourhood-level income quintile, when taken into account with all other predictors, was not statistically significantly associated with persistent high-cost use (p = 0.3). Several covariates (connection to a primary care provider, number of mental health conditions, number of comorbid conditions) interacted with ‘died during study period’. All interaction terms demonstrated increased risk for persistent high-cost use when study subjects died during the observation period. This occurred despite accounting for number of days alive in the study period when assigning costs and excluding long-term care residents in the modelling. Sensitivity analyses including and excluding patients who died during the study period did not differ in overall findings/conclusion.

A number of statistically significant two-way interactions occurred. Connection to a primary care provider was protective of persistent high-cost use in general, but particularly for individuals with multiple mental health conditions (p = 0.001). Unstable housing increased the probability of persistent high-cost use in general, but this effect was more pronounced with 2+ comorbid conditions (p = 0.02). Similarly, hospitalization increased the probability of persistent high-cost use, but particularly for those with two or more comorbidities (p = 0.01) (Table 2.3). The model showed evidence of good fit.
Table 2.3: Regression modelling, binary distribution, probit link, comparing persistent high-cost with not persistently high-cost use, mental health and addiction cohort, excluding long-term care residents, Saskatoon Health Region, FY2009–2015 (n = 124,497)

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Estimate</th>
<th>SE</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age &lt; 49 years</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age &gt;=50 years</td>
<td>0.14</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td>0.561</td>
</tr>
<tr>
<td>Male (reference)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>0.03</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>Mental health condition</td>
<td>&lt;0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anxiety disorders (reference)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mood/affective disorders</td>
<td>0.25</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>Schizophrenic disorders</td>
<td>0.94</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td>Substance-related disorders</td>
<td>0.13</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>Housing</td>
<td>&lt;0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stable housing (reference)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unstable housing</td>
<td>0.39</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td>Comorbid conditions</td>
<td>&lt;0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>None (reference)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>One</td>
<td>0.43</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>Two or more</td>
<td>0.92</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>Connection to a primary care provider</td>
<td>0.002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not connected (reference)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Well connected</td>
<td>0.09</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>Number of mental health conditions</td>
<td>&lt;0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>One (reference)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two or more</td>
<td>0.29</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>Hospitalization(s)</td>
<td>&lt;0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No (reference)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>0.54</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>Died during study period</td>
<td>&lt;0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No (reference)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>1.95</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>Connection to a primary care provider* Number of mental health conditions</td>
<td>0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Well connected, Two or more</td>
<td>0.12</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>Housing* Comorbid conditions</td>
<td>0.023</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unstable housing, One comorbid condition</td>
<td>-0.10</td>
<td>0.12</td>
<td></td>
</tr>
<tr>
<td>Unstable housing, Two or more comorbid conditions</td>
<td>-0.35</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td>Comorbid conditions* Hospitalization(s)</td>
<td>0.010</td>
<td></td>
<td></td>
</tr>
<tr>
<td>One comorbid, hospitalized</td>
<td>-0.05</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>Two or more, hospitalized</td>
<td>0.12</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>Comorbid conditions* Died during study period</td>
<td>&lt;0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>One, died</td>
<td>-0.23</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>Two or more, died</td>
<td>-0.38</td>
<td>0.05</td>
<td></td>
</tr>
</tbody>
</table>
### Number of mental health conditions* Died during study period

| Two or more, died | 0.33 | 0.05 | <0.001 |

### Connection to a primary care provider* Died during study period

| Well connected, died | -0.64 | 0.04 | <0.001 |

### Age* Mental health condition

| >=50 years, mood | -0.14 | 0.06 |
| >=50 years, schizophrenia | -0.73 | 0.10 |
| >=50 years, substance-related | -0.19 | 0.07 |

### Age* Sex

| >=50 years, female | 0.001 | 0.06 |

### Sex* Mental health condition

| Female, mood | -0.12 | 0.06 |
| Female, schizophrenia | 0.06 | 0.11 |
| Female, substance-related | -0.12 | 0.08 |

### Age* Sex* Mental health condition

| >=50 years, female, mood | 0.18 | 0.08 |
| >=50 years, female, schizophrenia | -0.30 | 0.14 |
| >=50 years, female, substance-related | 0.08 | 0.11 |

| Intercept | -3.09 | 0.05 |

Area under the curve = 0.90; Hosmer and Lemeshow goodness-of-fit: $\chi^2 = 11.1; df = 8; p = 0.2$ *Disorders of adult personality are included with mood/affective disorders due to small cell sizes (<5 when stratified)

All pairwise comparisons in interaction terms were adjusted for multiple comparisons (Bonferroni’s method; Supplemental material section 2.7.8).

### 2.5 Discussion

Average health care costs among study cohort members in FY 2009 were approximately $2,300; those among high-cost users were nearly $19,000. Despite making up less than 5% of the study population, persistent high-cost mental health and addiction clients in SHR accounted for nearly 35% of total SHR costs. In multivariable logistic regression, even after accounting for other potentially confounding factors, unstable housing was found to increase the probability of persistent high-cost use in the study population; a higher probability occurred if the individual had other underlying health conditions. A recent study in Ontario by Thavorn et al. similarly...
demonstrated, at an ecologic level, the association between residential instability and multi-morbidity in high costs (Thavorn et al. 2017).

Homelessness is known to be associated with increased health care utilization, poor health status, health inequities, and mental health conditions (Aldridge and Kelley 2015). The complexities of co-occurrence of mental illness, chronic medical conditions, and housing status need to be taken into account when conceptualizing and addressing multi-morbidity (Walker and Druss 2016). Individuals with complex MHA issues may cope with co-existing conditions (such as diabetes or heart disease) less well than individuals without mental illness (Prior et al. 2018). The positive effects of stable housing on health outcomes have been well documented—resulting in fewer communicable diseases, injuries, better chronic disease management/prevention, and improved psycho-social well-being (Fitzpatrick-Lewis et al. 2011).

Policy makers, concerned with health care cost containment, are reasonable to focus on the small groups of patients consuming the majority of resources. However, ‘high-cost users’ are not a homogenous group. Even when focusing on a particular patient group that is known to be high-cost, in this case mental health and addictions, and further refining the population of interest into persistently high-cost over many years, heterogeneity in the patient population still occurred. Despite this potentially heterogeneous study population, and taking into account other, potentially confounding, factors, such as health care utilization and demographic factors, unstable housing remained a significant risk factor for persistent high-cost use.

In terms of public health actions to address housing stability, a recent review of the evidence concluded policy makers should consider providing permanent supportive housing for homeless/disabled MHA patients (Rog et al. 2014). The provision of a stable home, particularly
for individuals with MHA and/or complex care needs, not only can decrease health care utilization/costs, but can improve health outcomes in a vulnerable population (Dieterich et al. 2017; Fitzpatrick-Lewis et al. 2011; Hwang et al. 2005; Poremski et al. 2016). It may also be reasonable to consider shifting resources, from costly acute care for mental health and addiction patients to providing better models of care in the community and connection to a primary care provider, particularly where multiple comorbid conditions and housing instability are present. A recent Canadian study demonstrated that community-based, coordinated access to care programs provide improved health outcomes, cost-effectively, among a cohort of homeless adults with mental health needs (Stergiopoulos et al. 2018). This points to the potential to identify mental health clients with unstable housing and comorbid conditions earlier and offer ‘Housing First’ or other supportive housing services, potentially preventing worsening of both their mental health and their other comorbidities and increasing their ability to cope (Chrystal et al. 2015; Poremski et al. 2016).

From published estimates (de Oliveira et al. 2017b, 2016c; Knapp et al. 2004), I know that schizophrenia and eating disorders are among the most expensive mental health and addiction diagnostic categories. In the current study, eating disorders are not delineated due to privacy considerations of small numbers in Saskatchewan (< 5 individuals per year); schizophrenic disorders were the most expensive diagnostic category.

As ‘high-cost use’ changes over time, even within the same individual, categories of persistent, sometimes, and never high-cost were created. Studies describing high-cost use as a snapshot in time may be combining potentially heterogeneous categories of individuals. As noted by Wodchis et al., one third of high-cost users in Ontario were found to be persistently high-cost in
the subsequent 2-year study follow-up (Wodchis et al. 2016); a similar proportion was found in the current study.

Not surprisingly, I found hospitalization increased the probability of persistent high-cost use. It is difficult to quantify ‘avoidable’ hospitalizations in administrative health data; however, a previous study in New Zealand estimated one in three hospitalizations was avoidable (Jackson and Tobias 2001).

I found that having a good connection to a primary care provider decreased the probability of being a high-cost user. Hospitalization is the costliest form of care in the health care system. Efforts to reduce persistent high-cost use could assess whether or not adequate community-based care could offset the costs associated with hospitalization(s).

Interestingly, despite accounting for number of days alive in the study period and excluding long-term care residents in the modelling, having died during the study period continued to be a significant risk factor for persistent high-cost use. Future studies could examine end-of-life costs, particularly if certain interventions can improve a patient’s experience while at the same time reducing costs.

This study has several limitations, some inherent to administrative health databases. By focusing on administrative health databases, I could not account for all potential confounders; for example, food insecurity (Tarasuk et al. 2015) has been found to be associated with high-cost use but not available for analysis. Community-based mental health service provision (counselling, treatment centres, others) may be associated with persistent high-cost use but data on these services were not available for analysis. My study focused on cost; however, not all health care costs were measured, such as home care, ambulatory care, out-of-pocket health care costs, such
as prescription medication costs covered by private insurance, travel costs (air transfers and ground ambulance) and all indirect costs (caregiving or lost wages). Misclassification likely occurred when categorizing individuals. For example, individuals with zero health care use and unstable housing would not be counted, underestimating the number of people with unstable housing. National data quality processes are in place for hospitalization and emergency department data; however, physician billing data can have diagnostic limitations. Despite this, reliability/validity has been found to be fairly good (Lix et al. 2013).

Results from this study demonstrate that efforts to reduce persistent high-cost use among a cohort of mental health and addiction clients should focus on multi-morbidity, connection to a primary care provider (particularly for those with more than one mental health condition), young patients with schizophrenia, and adequately addressing housing stability.
2.6 References


challenge to translate a utilization pattern into service provision. *Acta Psychiatrica Scandinavica.

Kerman, N., Sylvestre, J., Aubry, T., & Distasio, J. (2018). The effects of housing stability on
service use among homeless adults with mental illness in a randomized controlled trial of

Bulletin*, 30(2), 279–293.

based measure of the economic burden of mental illness in Canada. *Chronic Diseases in Canada,

Lix, L. M., Quail, J., Fadahunsi, O., & Teare, G. F. (2013). Predictive performance of
comorbidity measures in administrative databases for diabetes cohorts. *BMC Health Services

Poremski, D., Stergiopoulou, V., Braithwaite, E., Distasio, J., Nisenbaum, R., & Latimer, E.
(2016). Effects of housing first on employment and income of homeless individuals: results of a
/10.1176/appi.ps.201500002


2.7 Supplemental material

2.7.1 Cohort creation

For Chapters 2 and 3 the first step conducted was to create a mental health and addictions (MHA) cohort. This was achieved by pulling all records with MHA ICD-9 and ICD-10 codes, excluding dementia and dementia-related codes, (see section 2.7.6) from April 1 2003 to March 31 2015 in any of the following datasets: physician billing, hospitalization, long-term care and emergency department visits. Each unique scrambled study ID was then linked to the Person Health Registry System (PHRS) to obtain demographic data (age, sex, location, socio-economic indicators) at study baseline (April 1 2003). Individuals <18 years of age, not eligible for provincial health insurance and/or with >= 3 days gap in provincial health insurance coverage were excluded. Included individuals, identified by their unique study ID, constitute the study cohort. Once each unique study ID was identified, **all of their health care utilization data**, whether specific to mental health and addictions or not, were pulled from the administrative health databases (physician billing, hospitalization, long-term care and emergency department visits) for the duration of the study period. All study IDs were linked to a ‘verified death file’ created by the Saskatchewan Health Quality Council for research purposes to extract a date of death, where applicable.

A knowledgeable and experienced analyst, at my direction, conducted the aforementioned work. As per the Master Data Sharing Schedule, I am not permitted access to raw data files, and, the analyst understands and can use the cohort creation SQL macros that are pre-established, validated and modifiable to individual project specifications. Cohort creation pulls typically exceed 50 hours of computer run time.
Study cohort IDs were then linked to the PHRS to obtain demographic and socio-economic data at study baseline (April 1 2015) and, where applicable, to the ‘verified death file’ to obtain death dates.

2.7.2 Health care costing data

Micro-level, person-level health care costing methods were used to create provincial health care costing files. These provincial-level data files are available from April 1 2009 onwards and provide total health care costs, per service used, per year for every Saskatchewan resident with a valid health services card. The health care data used in creating this costing file are: physician billing, hospitalization, prescription drug, emergency department visits, home care visits and long term care residency. The file also provides the relative proportion each service constitutes per person per fiscal year. These files were created by a Saskatchewan-based health care economist accessing raw Saskatchewan administrative health data and using methods from Wodchis et al (developed in Ontario).

2.7.3 Other databases

Other provincial databases were accessed to extract data for study cohort members. These included: the ‘usual provider continuity’ index, the ‘shared client index’, the ‘street outreach’ database, the ‘Canadian Chronic Disease Surveillance System’ database and various other information sources (such as immunization data, communicable disease data and others).
All of the aforementioned database creation/pulls (2.7.1 to 2.7.3) were conducted by a knowledgeable and experienced analyst with permission to access raw data files, at my direction, using validated SAS code/macros. Each dataset was provided to me for linkage by unique study ID. **From this point forward, I conducted all analyses independently.**

### 2.7.4 Health care utilization and data manipulation

First, for Chapters 2 and 3, I limited the study population to Saskatoon Health Region residents only; the data required to calculate the variable ‘unstable housing’ was not available outside of Saskatoon. All individuals who died prior to April 1 2003 (study baseline) were excluded from further analysis. Once the study cohort(s) have been identified, and, all of their health care encounter and demographic data has been pulled for the relevant time period, I conducted months (years) of data manipulation to create the variables used in descriptive and analytic tables.

All health care utilization data (physician visits, emergency department visits, long-term care residency, and hospitalizations) were pulled from raw data files for study cohort members as separate files per databases in ‘long’ format (meaning: every study ID could have many records in up to four data files). I manipulated this data using ‘do loops’, ‘arrays’ and ‘transpose’ statements to create the variables I wanted to explore in the analyses.

#### 2.7.4.1 Hospital data

In the hospital data, I first excluded all hospital transfers as they are considered to be ‘one’ hospitalization and multiple transfers can occur during one episode of care. Transfers were
identified through a validated ‘episode of care’ variable created for research purposes by the Saskatchewan Health Quality Council. I calculated length of stay by subtracting the admission date from the discharge date. All hospitalizations for mental health (Chapters 2 and 3) were identified through the starting string value of the ICD-10 code ‘F’ in any of the diagnostic fields. All hospitalizations and lengths of stay were summed by fiscal year per person and transposed to ‘wide’ format; this created a data set with one record per study cohort member.

2.7.4.2 Physician billing data

In the physician data, I first excluded all ‘slush’ codes. This a group of specific physician codes that relate to out-of-province visits, telephone consults, insurance form completion, etc. and are typically excluded from research studies. Physician visits that occurred on the same day with the same physician were considered duplicate and excluded. Three digit ICD-9 diagnoses were categorized into ‘mental health’ or all others (see 2.7.6 for a list of MHA ICD codes). Physician visits were summed per fiscal year per person, using arrays, do loops and transpose statements, resulting in a data set with one record per study cohort member.

2.7.4.3 Emergency department

Emergency department data was fully complete for Chapters 2 and 3, as this study cohort focused on Saskatoon only. Emergency department visit data was manipulated similarly to physician and hospital data – summing the number of visits per fiscal year per study subject. Multiple visits by the same person on the same day, where the time stamp of arrival differed, were considered more than one visit.
2.7.4.4 Health care costing

Using the health care costing data files for the provincial population of Saskatchewan I conducted two analyses. First, I calculated the population threshold value for each fiscal year – this was the dollar amount associated with the 90th percentile in total health care costs. I calculated this value both for the entire fiscal year and per day based on the number of days of insurance coverage per fiscal year per study ID (coverage dates vary due to moving in and out of province and death). Second, I extracted total health care costs, by service used, per fiscal year, for each member of the study cohorts. Any study cohort member who exceeded the 90th percentile of total health care costs at the population level (population threshold value) was considered to be ‘high-cost’.

Persistent high-cost use (Chapters 2 and 3) was defined as any study cohort member exceeding the 90th percentile of total costs for each year under study (fiscal year 2009-2015). All other ‘high-cost’ definitions were episodic in nature (‘high-cost’ at one point in time, not necessarily persistently high-cost year after year).

Using the date of death, I calculated the number of days per fiscal year each study subject was alive and able to access health services. As I excluded all individuals with a $\geq 3$ day gap in health insurance coverage; for this reason, accounting for death was all that was needed to create a total cost per day variable. Therefore, I use a total cost per day population threshold value to determine the health care cost per day value exceeding the 90th percentile.
2.7.4.5 Other

Other provincial databases were accessed to extract data for study cohort members. These included: the ‘usual provider continuity’ index, the ‘shared client index’, the ‘street outreach’ database, the ‘Canadian Chronic Disease Surveillance System’ database and various other information sources (such as immunization data, communicable disease data and others). I extracted relevant data from each data source and linked the data by unique study ID. Many of these analyses were exploratory and ultimately not used in the Chapters included in this dissertation; mainly for data quality reasons.

Long-term care and home-care data were linked at the individual-level for each study cohort member. Total home-care visits by fiscal year were derived. Long-term care flags (Yes/No) per fiscal year were created for each study cohort member.

The variable ‘unstable housing’ (Chapters 2 and 3) was defined as a combination of individuals accessing street outreach nursing services and individuals with frequent address changes (as defined through address updates to the ‘Shared Client Index’, a database updated at each health service encounter in Saskatchewan).

To define comorbidities I used the algorithms developed by the Canadian Chronic Disease Surveillance System, and, where any study cohort member had >= 2 chronic diseases as defined by CCDSS algorithms, they were considered to be ‘multi-morbid’ (see 2.7.7).
2.7.5 Statistical modelling

In Chapter 2, I used a logistic regression with probit link approach to explore factors associated with both persistent and episodic high-cost use among a cohort of mental health and addiction clients in Saskatoon, Saskatchewan. As I was primarily interested in the binary outcome of high-cost health care use, employing logistic regression made the most sense. Odds ratios are inherently more intuitive to me; however, my Ontario collaborators were insistent that, in their experience, a probit link function outperformed the logit link. Interestingly, in conducting the analyses they were correct – when the link function was changed to probit from logit, the models showed evidence of better fit and less heteroskedasticy. Model building efforts were a combination of art and science. Directed acyclic graphs of variable association with the outcome were created for each study cohort (Chapters 2, 3 and 4). This included assessments of confounding (review of model co-efficients: did inclusion of the potentially confounding variable change the coefficient substantially, defined at >=20% change in coefficient estimates), and, creating interaction terms that made biological sense and assessing their statistical significance. In reality, I assessed all potential 2-way interactions and many 3-way interactions. Only those interactions that a) improved model fit, and, b) were identified a priori by clinical collaborators were ultimately included in final models. In model building, I used the Akaike Information Criterion (AIC) to assess the incremental gains/losses in the AIC value with various iterations of models. Where the change in AIC values from model to model varied slightly (for example, differences of >=1.0 I chose the simplest most parsimonious model. All variables were entered into the models, with the exception of Chapter 3 where effect selection in the LASSO framework was conducted prior to constructing classification trees; all effect selection was done by me, not forward/backward selection procedures in the statistical software. Model diagnostics
were performed with creation of each of the models to assess the assumptions of linearity and independence. For each final model I conducted a sensitivity analysis of forward selection to compare predictors I used versus the ones SAS ended up including. With the exception of confounding variables and discovery of important interaction terms, the models created by SAS were similar to the ones I created.
2.7.6 International Classification of Diseases (ICD-9 and ICD-10 CA) codes by MHA condition

<table>
<thead>
<tr>
<th>Condition*</th>
<th>ICD-9</th>
<th>ICD-10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Substance related disorders</td>
<td>291, 292; 303-305</td>
<td>F10-F19; F55</td>
</tr>
<tr>
<td>Schizophrenia, schizotypal, delusional and other non-mood psychotic disorders</td>
<td>295, 297, 298</td>
<td>F20-F25; F28, F29</td>
</tr>
<tr>
<td>Mood/affective disorders</td>
<td>296, 311</td>
<td>F30-F34; F39</td>
</tr>
<tr>
<td>Anxiety disorders</td>
<td>300, 308, 309</td>
<td>F40-F48; F93</td>
</tr>
<tr>
<td>Disorders of adult personality</td>
<td>301</td>
<td>F60-F69</td>
</tr>
</tbody>
</table>

* Individuals >=18 years, with one or more physician claims with the relevant ICD-9 diagnostic code, or, one ICD-10-CA hospital discharge in any diagnosis field(s)

2.7.7 Chronic disease case definitions: Canadian Chronic Disease Surveillance System

<table>
<thead>
<tr>
<th>Condition</th>
<th>Case definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asthma</td>
<td>A person with:</td>
</tr>
<tr>
<td></td>
<td>• One or more hospital discharge records with a discharge diagnosis of asthma in any diagnostic field. ICD-10-CA: J45, J46 AND/OR • Two or more physician service claims with an asthma diagnosis within two years. The index date is assigned the date on the latest physician claim. ICD-9: 493 If an individual meets both criteria, the index date of diagnosis is the earliest date of either the hospitalization or the physician service claim. No age restriction.</td>
</tr>
<tr>
<td>Chronic obstructive pulmonary disorder</td>
<td>A person with:</td>
</tr>
<tr>
<td></td>
<td>• One or more hospital discharge records with a discharge diagnosis of COPD in any diagnostic field. ICD-10-CA: J41-J44 AND/OR • One or more physician service claims with a COPD diagnosis. ICD-9: 491, 492, 496 If an individual meets both criteria, the index date of diagnosis is the earliest date of either the hospitalization or the physician service claim. Exclude age under 35.</td>
</tr>
<tr>
<td>Diabetes</td>
<td>A person with:</td>
</tr>
<tr>
<td></td>
<td>• One or more hospital discharge records with a discharge diagnosis of COPD in any diagnostic field. ICD-10-CA: E10, E11, E12, E14, I13 AND/OR • Two or more physician service claims with a diabetes diagnosis within two years. The index date is assigned the date on the latest physician claim.</td>
</tr>
<tr>
<td>Condition</td>
<td>Definition</td>
</tr>
<tr>
<td>------------------------------------------</td>
<td>-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
</tbody>
</table>
| Coronary artery disease                  | Individuals aged 20 and over with: One or more hospital discharge records with a discharge diagnosis of CAD in any diagnostic field. ICD-10-CA: I20, I21, I22, I23, I24, I25  
   AND/OR  
   • One or more hospital discharge records with an intervention code of percutaneous coronary intervention (PCI) or coronary artery bypass grafting (CABG) in any intervention field. CCI: 1I50, 1I57GQxx, 1I76  
   AND/OR  
   • Two or more physician service claims with a CAD diagnosis within 1 year. ICD-9: 410-414 If an individual meets both criteria, the index date of diagnosis is the earliest date of either the hospitalization or the physician service claim. |
| Congestive heart failure                 | A person with: One or more hospital discharge records with a discharge diagnosis of CHF in any diagnostic field. ICD-10-CA: I50.0, I50.1, I50.9  
   AND/OR  
   • Two or more physician service claims within one year with a CHF diagnosis. ICD-9: 428 If an individual meets both criteria, the index date of diagnosis is the earliest date of either the hospitalization or the physician service claim. |

Source: Canadian Chronic Disease Surveillance System (CCDSS)
2.7.8 Probability estimates of interaction terms, regression model mental health and addictions cohort, excluding long-term care residents, Saskatoon Health Region, FY2009-2015 (n = 124,497), Bonferroni’s pairwise comparisons

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Estimate</th>
<th>95% CI</th>
<th>Group*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male, &lt; 49 years, anxiety</td>
<td>0.001</td>
<td>0.0007-0.001</td>
<td>C</td>
</tr>
<tr>
<td>Male, &lt; 49 years, mood/affective disorders</td>
<td>0.002</td>
<td>0.002-0.003</td>
<td>B</td>
</tr>
<tr>
<td>Male, &lt; 49 years, schizophrenia</td>
<td>0.016</td>
<td>0.011-0.022</td>
<td>A</td>
</tr>
<tr>
<td>Male, &lt; 49 years, substance-related disorders</td>
<td>0.002</td>
<td>0.001-0.002</td>
<td>B, C</td>
</tr>
<tr>
<td>Female, &lt; 49 years, anxiety</td>
<td>0.001</td>
<td>0.0008-0.001</td>
<td>C</td>
</tr>
<tr>
<td>Female, &lt; 49 years, mood/affective disorders</td>
<td>0.002</td>
<td>0.001-0.002</td>
<td>B</td>
</tr>
<tr>
<td>Female, &lt; 49 years, schizophrenia</td>
<td>0.019</td>
<td>0.013-0.029</td>
<td>A</td>
</tr>
<tr>
<td>Female, &lt; 49 years, substance-related disorders</td>
<td>0.001</td>
<td>0.001-0.002</td>
<td>B, C</td>
</tr>
<tr>
<td>Male, ≥ 50 years, anxiety</td>
<td>0.002</td>
<td>0.001-0.002</td>
<td>B, C</td>
</tr>
<tr>
<td>Male, ≥ 50 years, mood/affective disorders</td>
<td>0.003</td>
<td>0.002-0.004</td>
<td>A</td>
</tr>
<tr>
<td>Male, ≥ 50 years, schizophrenia</td>
<td>0.002</td>
<td>0.001-0.002</td>
<td>A</td>
</tr>
<tr>
<td>Male, ≥ 50 years, substance-related disorders</td>
<td>0.001</td>
<td>0.001-0.002</td>
<td>C</td>
</tr>
<tr>
<td>Female, ≥ 50 years, anxiety</td>
<td>0.002</td>
<td>0.001-0.002</td>
<td>B</td>
</tr>
<tr>
<td>Female, ≥ 50 years, schizophrenia</td>
<td>0.001</td>
<td>0.001-0.002</td>
<td>B</td>
</tr>
<tr>
<td>Not connected to primary care provider*One MH condition</td>
<td>0.001</td>
<td>0.0007-0.001</td>
<td>A</td>
</tr>
<tr>
<td>Connected to primary care provider*One MH condition</td>
<td>0.001</td>
<td>0.001-0.002</td>
<td>B</td>
</tr>
<tr>
<td>Not connected to primary care provider*Two or more mental health conditions</td>
<td>0.003</td>
<td>0.002-0.003</td>
<td>C</td>
</tr>
<tr>
<td>Connected to primary care provider*Two or more mental health conditions</td>
<td>0.005</td>
<td>0.004-0.006</td>
<td>D</td>
</tr>
<tr>
<td>Unstable housing, no comorbid conditions</td>
<td>0.003</td>
<td>0.002-0.005</td>
<td>B</td>
</tr>
<tr>
<td>Unstable housing, one comorbid condition</td>
<td>0.009</td>
<td>0.005-0.015</td>
<td>A, B</td>
</tr>
<tr>
<td>Unstable housing, two or more comorbid conditions</td>
<td>0.016</td>
<td>0.009-0.029</td>
<td>A</td>
</tr>
<tr>
<td>Stable housing, no comorbid conditions</td>
<td>0.001</td>
<td>0.007-0.001</td>
<td>C</td>
</tr>
<tr>
<td>Stable housing, one comorbid condition</td>
<td>0.004</td>
<td>0.003-0.005</td>
<td>B</td>
</tr>
<tr>
<td>Stable housing, two or more comorbid conditions</td>
<td>0.016</td>
<td>0.009-0.029</td>
<td>A</td>
</tr>
<tr>
<td>Hospitalized, no comorbid conditions</td>
<td>0.005</td>
<td>0.002-0.007</td>
<td>D</td>
</tr>
<tr>
<td>Hospitalized, one comorbid condition</td>
<td>0.015</td>
<td>0.012-0.018</td>
<td>B</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------------------</td>
<td>---------</td>
<td>-------------</td>
<td>---</td>
</tr>
<tr>
<td>Hospitalized, two or more comorbid conditions</td>
<td>0.065</td>
<td>0.054-0.077</td>
<td>A</td>
</tr>
<tr>
<td>Not hospitalized, no comorbid conditions</td>
<td>0.001</td>
<td>0.0007-0.001</td>
<td>E</td>
</tr>
<tr>
<td>Not hospitalized, one comorbid condition</td>
<td>0.004</td>
<td>0.003-0.005</td>
<td>C</td>
</tr>
<tr>
<td>Not hospitalized, two or more comorbid conditions</td>
<td>0.015</td>
<td>0.011-0.018</td>
<td>B</td>
</tr>
</tbody>
</table>

*Adjusted for multiple comparisons (Bonferroni method); Estimates with the same letter are not significantly different.
2.7.9 Co-author affiliations and contributions

Maureen Anderson MSc, PhD (candidate)\textsuperscript{1,2*}, Crawford W Revie MSc, PhD\textsuperscript{1}, Jacqueline M Quail MSc, PhD\textsuperscript{2,6}, Walter Wodchis MA, MAE, PhD\textsuperscript{3,4}, Claire de Oliveira MA, PhD\textsuperscript{3,4,5}, Meriç Osman MA\textsuperscript{2}, Marilyn Baetz MD\textsuperscript{6,7}, J McClure DVM, MS\textsuperscript{1}, Henrik Stryhn MSc, PhD\textsuperscript{1}, David Buckeridge MD, MSc, PhD\textsuperscript{8}, Cordell Neudorf MD, MHSc\textsuperscript{6,7}

\textsuperscript{1} Department of Health Management, Atlantic Veterinary College, University of Prince Edward Island, 550 University Avenue, Charlottetown, PE C1A 4P3
\textsuperscript{2} Saskatchewan Health Quality Council, Atrium Building, Innovation Place, 241 – 111 Research Drive, Saskatoon, SK S7N 3R2
\textsuperscript{3} Institute of Health Policy, Management and Evaluation, University of Toronto, Health Sciences Building 155 College Street, Suite 425, Toronto, ON M5T 3M6
\textsuperscript{4} Institute for Clinical Evaluative Sciences, 2075 Bayview Ave, Toronto, ON M4N 3M5
\textsuperscript{5} Centre for Addiction and Mental Health, 1001 Queen Street West, Toronto, Ontario M6J 1H4
\textsuperscript{6} College of Medicine, University of Saskatchewan, 107 Wiggins Road, Saskatoon, SK S7N 5E5
\textsuperscript{7} Saskatoon Health Region, #101–310 Idylwyld Drive North, Saskatoon, SK, S7L 0Z2
\textsuperscript{8} Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Purvis Hall 1020 Pine Avenue West, Montreal, QC, Canada H3A 1A2

MA – Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Software, Validation, Visualization, Writing – original draft preparation, Final approval and Agreement to be accountable

CR – Conceptualization, Methodology, Visualization, Supervision, Writing – original draft preparation, Writing – review and editing, Final approval and Agreement to be accountable

JQ – Conceptualization, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Writing – review and editing, Final approval and Agreement to be accountable

WW – Conceptualization, Funding Acquisition, Methodology, Supervision, Writing – review and editing, Final approval and Agreement to be accountable

CO – Conceptualization, Funding Acquisition, Methodology, Supervision, Writing – review and editing, Final approval and Agreement to be accountable

MO – Data Curation, Formal Analysis, Investigation, Methodology, Software, Validation, Writing – review and editing, Final approval and Agreement to be accountable

MB – Methodology, Validation, Writing – review and editing, Final approval and Agreement to be accountable

JM – Conceptualization, Methodology, Supervision, Final approval and Agreement to be accountable
HS – Conceptualization, Methodology, Supervision, Validation, Writing – review and editing, Final approval and Agreement to be accountable

DB – Conceptualization, Validation, Supervision, Final approval and Agreement to be accountable

CN – Conceptualization, Funding Acquisition, Resources, Supervision, Writing – review and editing, Final approval and Agreement to be accountable
3. USING CLASSIFICATION AND REGRESSION TREE ANALYSIS TO PREDICT FUTURE HIGH-COST HEALTHCARE USE: RESULTS FROM A RETROSPECTIVE POPULATION-BASED COHORT STUDY OF MENTAL HEALTH AND ADDICTION CLIENTS IN SASKATOON, SASKATCHEWAN


3.1 Abstract

Objective: To predict one-time and persistent high-cost healthcare use.

Methods: I conducted a retrospective population-based cohort study in Saskatoon, Saskatchewan. Classification and regression tree analyses (CART) were used for prediction.

Results: A total of 124,497 cohort members were identified. The majority (~65%) were ‘never high-cost’. Persistent high-cost users (n=2,352; 2%) accounted for 15% of costs.

The most important variables for predicting one-time high-use were: 1) number of physician visits prior to high-use; 2) healthcare costs prior; 3) number of days in hospital prior; 4) number of comorbid conditions. For persistent high-use: 1) number of physician visits prior; 2) the primary mental health condition of the patient (depression vs. other).

Conclusion: The predictive models take advantage of information readily available in Canadian provincial administrative healthcare databases. Predictive algorithms can provide actionable data to prevent future high-use. The visual nature of decision-trees facilitates model understanding.

Max 150 words: currently 149

Key words: high-cost users; mental health and addictions; predictive modelling
3.2 Introduction

Evidence has long demonstrated that a small proportion of the population (≤10%), commonly referred to as ‘high-cost users’, accounts for the majority (50-70%) of total healthcare spending (Densen, Shapiro, & Einhorn, 1959; N. Roos, Burchill, & Carriere, 2003; N. P. Roos, Shapiro, & Tate, 1989; Stuart & Weinrich, 1998). High-cost users are a heterogeneous group (de Oliveira, Cheng, Vigod, Rehm, & Kurdyak, 2016; Taube, Goldman, Burns, & Kessler, 1988). As early as 1988, Taube et al. demonstrated that individuals with mental disorders comprised 9% of high utilization patients (i.e. those with ≥25 visits) but accounted for nearly half (49.2%) of total outpatient expense (Taube et al., 1988). In a Medicaid population study of high-cost users, Buck et al. (2003) showed that people using mental health and substance abuse services constituted 11% of all enrollees but accounted for approximately 30% of all high-cost users (Buck, Teich, & Miller, 2003). Hunter et al. (2015) found that nearly half of high-cost users had a mental health condition (Hunter, Yoon, Blonigen, Asch, & Zulman, 2015).

Previously, I described persistent high-cost users with mental health and addiction (MHA) issues and quantified, taking into account other potentially confounding factors, factors that were associated with persistent high-cost use (Anderson et al., 2018). The current study aims to build upon these findings, and predict, with an adequate degree of accuracy, future high-cost healthcare use.

The majority of previous studies aimed at predicting future high-cost healthcare were conducted in the United States. Wrathall and Belnap, at Intermountain Healthcare in Utah, determined a logistic model containing clinical characteristics and socio-demographic characteristics performed best in predicting future high-cost use (Wrathall & Belnap, 2017). In a Florida
Medicaid study, Robst found that institutional living and prior mental health and addiction healthcare costs were most predictive of future persistent high-cost use among a cohort of mental health clients (Robst, 2015). Fleishman and Cohen, using American survey data, found that the best-fit models were those which took into account the number of comorbid chronic conditions in predicting future healthcare costs (Fleishman & Cohen, 2010). Lu and colleagues, in Virginia, similarly found a count of chronic conditions, coupled with congestive heart failure and demographic information, could best predict future high-cost use (Lu, 2015). In 2016, Chang and colleagues found that a model taking into account only medication costs performed better than models which accounted for diagnostic groupings, medication and prior healthcare costs when predicting consistent high-cost healthcare users (Chang et al., 2016).

I found one previous Canadian study aiming to construct a predictive model for high-cost healthcare use; an additional Canadian study aimed at predicting mental health high-cost patients is currently in press (in press: de Oliveria et al). Chechulin et al, developed logistic models to predict persons at risk of future high-cost use in Ontario (excluding individuals < 5 years of age; high-cost use defined as individuals in the top 5% of costs in one year). The authors found the strongest predictors in their model were age, chronic health conditions, measures of healthcare utilization (particularly, mental health inpatient stay), long term care residency and dialysis. In addition, neighbourhood level measurements of material deprivation increased the odds of high-cost use (Chechulin, Nazerian, Rais, & Malikov, 2014). The authors posit the model could be used for predicting, but, there is little evidence to suggest it has been used beyond the classical epidemiological sense. Summary statistics for these models are provided (Table 2).

Applying innovative analytical techniques, including prediction, to both common (for example, administrative health data) and less common (for example, social media data) data sources may
provide important evidence to inform health policy and intervention. The use of predictive algorithms in marketing is not new; the use of the same in health care has, arguably, reached less than its full potential (Yeung, 2018). It is concerning that private industry uses ‘big data’ and predictive algorithms from multiple sources to target advertisements with such regularity it is commonplace; but the same cannot be said for models geared to improve the health of the population.

Given the differences in descriptive epidemiology between them, I hypothesized that a stratified approach to prediction, namely, modelling persistent high-cost use separately from one-time high-cost use, would outperform a common model. In addition, the relative dearth of predictive modelling studies for high-cost healthcare users in the Canadian context is problematic. Canadian administrative health data is population-based; as opposed to American studies which are limited to specific populations (Medicaid, Veteran’s, and Health Management Organizations) and may not be representative of the population as a whole. Our work is important to both delineate findings that may be different from previous work specific to a Canadian context, to develop separate models for one-time and persistent high-cost use, and/or confirm similar findings.

The objective of the study was to develop, from the perspective of a provincial health insurer, a useful predictive model, using socio-demographic, clinical and healthcare utilization variables prior to the outcome of high-cost use to identify individuals at risk of becoming high-cost in the future, among a cohort of MHA clients in Saskatoon, Saskatchewan. I aim to predict both 1) persistent high-cost use and 2) one-time high-cost healthcare use, and, take into account the variety of factors that have previously been found to be predictive of future high-costs.
Given the study aim of predicting future high-cost use at the individual level and the difficulty in obtaining comprehensive individual-level data on socio-demographic predictors that have previously been shown to be associated with high-cost use, such as, unstable housing, food insecurity, and obesity (Anderson et al., 2018; Thavorn et al., 2017; Wrathall & Belnap, 2017) I opted to employ Classification and Regression Tree (CART) analyses.

Classification and regression tree analyses are not new; however, they are increasingly being used as a technique for modelling large, complex datasets. Even 15 years ago, researchers lacked the computational power to perform CART analysis, particularly in large datasets with many potential predictors. In 2019, I can construct and run these models with relative ease. The added benefit of the CART approach is its visual output where prediction becomes straightforward, even for individuals not well versed in the technique – a new input transverses over the tree starting at the root node. In addition, the method provides a more easily understood output for decision-makers by way of a visual identification of potential categorical breaks in continuous predictors of future high cost use. CART is also useful for identifying variables of relative importance; important when imputing multiple, seemingly similar, healthcare utilization variables (Breiman, Friedman, Stone, & Olshen, 1984).

Creating a model with a high degree of specificity could provide the evidence for public health action to mitigate high-cost use. In contrast to reactive intensive case management strategies, a well-calibrated and validated predictive model could provide evidence for strategies aimed at preventing high-cost use from occurring in the first place.
3.3 Methods

Saskatchewan (population ~1.2 million) is a Canadian province with a central provincial health insurer. Healthcare data are captured for the provincial population with the exception of residents covered under the federal government (~1%), specifically, military personnel, federal penitentiary inmates, and, prior to April 1 2013, members of the Royal Canadian Mounted Police (RCMP) (Downey, Stang, Beck, Osei, & Nichol, 2006). Data available for this study was limited to the largest health region in the province, Saskatoon Health Region (n = ~360,000; an organizational entity that no longer exists, Saskatchewan’s 12 regional health authorities were amalgamated into one effective December 2017). As the study population was limited to one specific geographic area in Canada, it is unknown to what degree the results are generalizable to other populations, however, one would expect reasonably generalizable to other provincial/territorial jurisdictions in Canada.

**Databases:** Detailed descriptions of Saskatchewan Ministry of Health databases are available elsewhere (Downey et al., 2006). In brief, demographic characteristics, location of residence and insurance coverage were extracted from the *Personal Health Registration System (PHRS)*. Hospital data were extracted from the Canadian Institute for Health Information (CIHI) validated *Discharge Abstract Database (DAD)* for the province of Saskatchewan (which includes all acute care and psychiatric hospitalizations in the province). The International Classification of Diseases (ICD), 10th revision, Canadian Version (ICD-10-CA) was used in the DAD to record up to 25 diagnoses at admission, including the diagnosis most responsible for the hospitalization at discharge. Data on physician services are contained in the *Medical Services Database.*

Physicians paid on a fee-for-service basis submit billing claims to the provincial health ministry; a single diagnosis using a three-digit ICD-9 code is recorded on each claim. Salaried physicians
are required to submit billing claims for administrative purposes (shadow billing); however, claims from salaried physicians may involve under-reporting resulting in fewer physician claims in the data than in practice (Saskatchewan Ministry of Health, 2017). Residents of long-term care facilities were defined according to the provincial Resident Assessment Instrument-Minimum Dataset (RAI-MDS) for long-term care facilities (RAI-LTC 2.0). The following datasets were available for Saskatoon Health Region only: emergency department data recorded in the National Ambulatory Care Reporting System (NACRS), the Shared Client Index (SCI) an identity management system used to manage patient contact information, and the Population and Public Health, Saskatoon Health Region, Street Outreach Database. The latter two databases permitted an administrative database definition of ‘unstable housing’ – any individual in the street outreach database and any individual with >= 4 address changes within a 12-month period from the SCI. Data were anonymized and linked at the individual level using a unique non-identifiable health services number generated by eHealth Saskatchewan.

**Study Design:** I conducted a retrospective population-based cohort study using administrative healthcare data. The study population included any individual with continuous provincial insurance coverage, resident of Saskatoon Health Region, >=18 years old, and alive as of April 1st, 2009 (study baseline) with at least one mental health and addiction-related ICD-9 or ICD-10-CA diagnosis between April 1st, 2003 and March 31st, 2015 with the exception of dementia and dementia-related codes. This sensitive case definition facilitated casting a wide net of all potential MHA patients in the study population; patients were then further refined into diagnostic groups based on severity of illness (anxiety/mood disorders to schizophrenic/delusional) as best can be ascertained by administrative health data where additional information, such as, presence or absence of suicidal ideation is not available.
Outcome and covariates: Outcome variables for this study were explored in separate models: 1) High-cost healthcare use in one year (≥90th percentile of costs in FY2010) and 2) Persistent high-cost healthcare use (similarly, ≥90th percentile of costs, but continuously for fiscal years 2009 to 2015). All covariates were defined relative to the study baseline (FY2009), with the exception of healthcare utilization and costing data. Healthcare utilization variables were defined as the total number of physician visits, psychiatrist visits, emergency department visits, hospitalizations and hospital length of stay, and, for ease of interpretation of the models, at two time points: 1) two years prior to high-cost use and 2) one year prior (Table 1). Healthcare utilization variables were assigned to their fiscal year of occurrence by admission date. In predictive modelling, select comorbid conditions were defined at the individual-level using available, validated administrative case definitions: asthma, diabetes, COPD, ischemic heart disease and heart failure (Feely, Lix, & Reimer, 2017). Other comorbidity indices, such as the Elixhauser or Charlson Index, were not employed; in Saskatchewan they are available for hospitalized patients only. A categorical variable to indicate the total number of selected comorbid conditions identified was created (0, 1 and 2+). Healthcare utilization data were analysed from FY2003 to 2015 healthcare costing data was analysed for FY2009 to 2015. Healthcare costs were adjusted for inflation to the year 2015.

Healthcare costs: Total healthcare costs were calculated for unique individuals for fiscal years 2009-2015. Healthcare costs were available for hospitalizations (both in-patient and day surgery for all conditions including mental health, excluding transfers), physician visits (both in-patient and office visits), emergency department visits, and prescription drugs. Costing variables were assigned to their fiscal year of occurrence from FY2009 to 2015; models predicting one year of high-cost use included the total healthcare costs incurred the year prior. High-cost status among
cohort members was determined for each year of the study period using population proportional thresholds (Wodchis, Austin, & Henry, 2016). Population threshold values refer to the dollar amount associated with the 90th percentile of health care costs per fiscal year of total health care costs for the entire provincial population, not just those under study. Once this dollar amount is calculated, then, within each study cohort(s), any individual exceeding this dollar amount, per fiscal year, will be considered ‘high-cost’. To account for death during the study period, each subject’s total healthcare costs were divided by total number of days observed in the study. Deaths were included in the study population.

**Statistical Analyses:** Effect selection in the framework of a median quantile regression with LASSO selection method (Tibshirani, 1996) was used to limit the number of covariates to those contributing most to outcome measures by choosing the model with the smallest Akaike’s information criterion (AIC) (Akaike, 1973). These variables were then used in classification and regression tree analyses to predict future high-cost healthcare use (Breiman et al., 1984). Sensitivity analyses were performed to compare results from decisions trees prior to effect selection and following.

As a first step, a full tree was built using all of the input data. Further refinements were made through pruning using cost-complexity methods; pruning reduces the complexity of the final model, reduces overfitting and improving predictive accuracy. Cost-complexity creates an initial tree with a root alone and many sub-trees; the best tree is then chosen from these through the use of a training set (cross-validation). Each terminal node was set to include a minimum of 5 individuals to comply with health data privacy regulations. Splitting criteria were based on improvements in entropy. Gini impurity scores, in addition to sensitivity/specificity and area under the curve values, were used to evaluate discriminatory and calibration gains in the final
tree (Breiman et al., 1984). A random sampling approach was used to partition all modelling data into training (60%) and validation (40%) datasets (Dobbin & Simon, 2011). Models were fit on training data only; prediction results were obtained for the validation dataset.

As the goal of this study was to create a model that predicts future high-cost users, and LTC use is a known high-cost use of healthcare, it was considered reasonable to exclude these study subjects. All long-term care (LTC) residents (n=5,435; 4.2% of cohort) were excluded from the modelling. All residents outside of the former Saskatoon Health Region area were excluded from the study cohort as all data was not available for the provincial population. Sensitivity analyses were conducted including and excluding deaths (n = 5,379; 4.1% of the cohort). Where deaths were included, the variables considered ‘most important’ for predicting future high-cost varied slightly, but, overall model interpretations remained the same.

Readers will note the population and outcome(s) under study described in this Chapter are similar to those in Chapter 2. A short explanation of differences, however, is warranted. In Chapter 2, all descriptive analyses included long-term care residents (n = 129,932); as the research progressed on Chapter 2, I came to understand a large proportion of long-term care residents were persistent high-cost users. Therefore, the current Chapter, both descriptive (Table 3.1) and regression analyses, focus solely on a population excluding long-term care residents from the start (n = 124,497). Readers will note that in Chapter 2 descriptive tables (Tables 2.1 and 2.2), then, persistent high-cost users number 6,455 but in the current chapter number 2,396. Rationale for this difference:

Chapter 2: Probit models (outcome: n = 4,107)

\[
\text{Model outcome} = \text{persistent high-cost users} - \text{long-term care residents}
\]

\[
6,455 - 2,348 = 4,107
\]
Chapter 3: CART models (outcome: n = 2,396)

Model outcome = persistent high-cost users – long-term care residents/died

\[ 6,455 - 4,059 = 2,396 \]

All analyses were conducted using SAS© Enterprise Guide version 7.1 (SAS Institute Inc., 2017).

The study proposal underwent ethical review and approval by the University of Saskatchewan Biomedical Research Ethics Board and the University of Prince Edward Island Research Ethics Board for research involving humans.

### 3.4 Results

A total of 124,497 unique individuals eligible for provincial health insurance, resident of Saskatoon Health Region (SHR), excluding long-term care residents, >=18 years of age and alive at study baseline were identified (Figure 1). Of these, the majority (~65%) were ‘never high-cost’ over the study period. Cohort characteristics are described (Table 3.1).

**Descriptive analyses: Predicting one year high-cost use**

Using 90th percentile population threshold values above which a study cohort member would be considered ‘high-cost’, approximately 12% of the study cohort was a high-cost user in FY2010. Of those high-cost in FY2010, approximately 30% continued to be high-cost in the next two years (n=4,850; 33%). All available variables (healthcare utilization, socio-demographic and healthcare costs) were entered into effect selection models aiming to determine variables associated with high-cost use in FY2010; using LASSO effect selection the following variables dropped out and not included in subsequent CART modelling: urban/rural location,
neighbourhood income quintile, mental health-specific hospital lengths of stay (FY2008), and emergency department visits (FY2008). All remaining variables were used in Classification and Regression Tree (CART) analyses predicting future high-cost use.

**Descriptive analyses: Predicting persistent high-cost use**

Persistent high-cost users, defined as top 10% of costs for the duration of the costing period (FY2009 to 2015) were a small component of the study population (n = 2,352; 2%) but accounted for 15% total healthcare costs. Persistent high-cost user characteristics are described (Table 3.1).

Psychiatrist visits, hospitalizations and emergency department visits were of similar magnitude between persistent and not persistent high-cost healthcare users; however, mean physician visits showed evidence of an increase. To explore this potential relationship further, the average number of physician visits was plotted over time. The number of physician visits by year, including 95% confidence intervals, showed a marked increase for persistent high-cost users over time, but, not for non-persistent high-cost users.

**Model 1: Predicting high-cost use in one year**

Using CART, the most important variables for predicting episodic high-cost use in the study cohort were (in order of importance): 1) the number of physician visits (one year prior to high-cost use), 2) total health care costs (one year prior), 3) number of physician visits (two years prior) and, 4) comorbid conditions and 5) length of stay in hospital (one year prior), (Figure 3.2). The validation model (n = 49,875) demonstrated evidence of good discrimination (AUC=0.76), calibration (Gini impurity = 18%), low sensitivity (19%) and high specificity (99%) (Breiman et al., 1984).
Model 2: Predicting persistent high-cost use

The most important variables for predicting persistent high-cost use in the study cohort were (in order of importance): 1) the number of physician visits (one year prior to high-cost use), 2) the number of physician visits (two years prior), and, 3) primary mental health condition (Figure 3.3).

The model showed evidence of good discrimination (AUC=0.82), fair calibration (Gini impurity = 4%), low sensitivity (2%) and high specificity (99%) (Breiman et al., 1984).
Figure 3.1: Study inclusion/exclusion criteria

Eligible study population: all unique persons with at least one mental health and addictions ICD-code, excluding dementia, in administrative health databases (physician billing, hospitalization, and/or emergency department) April 1st, 2003 to March 31st, 2015

n = 451,674

Exclude residents outside of Saskatoon Health Region (study population) and alive as of April 1st 2009 (study baseline)

n = 162,566

Exclude deaths prior to April 1st, 2009

n = 138,257

Exclude those ineligible for provincial health insurance and/or without continuous health coverage (>=3 day gap)

n = 129,932

Exclude long-term care residents

n = 124,497

High-cost users FY2010, exclude deaths prior to April 1 2010

n = 14,848

Persistent high-cost users, exclude deaths

n = 2,396
Table 3.1 Mental health and addiction high-cost cases, Saskatoon, Saskatchewan, demographics and health care utilization by cost outcome (Study cohort n = 124,497)

<table>
<thead>
<tr>
<th></th>
<th>High-cost FY2010 (n = 14,848)</th>
<th>Persistent high-cost (n = 2,396)</th>
<th>Never high-cost (n=71,821)</th>
<th>Study population overall (n=124,497)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-64 yrs</td>
<td>10,436 (70.3%)</td>
<td>1,738 (72.5%)</td>
<td>67,376 (93.8%)</td>
<td>108,299 (87.0%)</td>
</tr>
<tr>
<td>65+ yrs</td>
<td>4,412 (29.7%)</td>
<td>658 (27.5%)</td>
<td>4,445 (6.2%)</td>
<td>16,198 (13.0%)</td>
</tr>
<tr>
<td><strong>Sex</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>5,528 (37.2%)</td>
<td>1,015 (42.4%)</td>
<td>34,231 (47.7%)</td>
<td>53,570 (43.0%)</td>
</tr>
<tr>
<td>Female</td>
<td>9,320 (62.8%)</td>
<td>1,381 (57.6%)</td>
<td>37,590 (52.3%)</td>
<td>70,927 (57.0%)</td>
</tr>
<tr>
<td><strong>Geographic location</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Urban</td>
<td>12,359 (83.2%)</td>
<td>2,070 (86.4%)</td>
<td>60,657 (84.5%)</td>
<td>104,469 (83.9%)</td>
</tr>
<tr>
<td>Rural</td>
<td>2,341 (15.8%)</td>
<td>311 (12.9%)</td>
<td>10,283 (14.3%)</td>
<td>18,580 (14.9%)</td>
</tr>
<tr>
<td>Missing</td>
<td>148 (1%)</td>
<td>15 (0.6%)</td>
<td>881 (1.2%)</td>
<td>1,448 (1.2%)</td>
</tr>
<tr>
<td>Neighbourhood income</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 (least affluent)</td>
<td>3,277 (22.1%)</td>
<td>558 (23.3%)</td>
<td>11,327 (18.2%)</td>
<td>22,006 (19.5%)</td>
</tr>
<tr>
<td>2</td>
<td>2,571 (17.3%)</td>
<td>411 (17.2%)</td>
<td>10,855 (17.4%)</td>
<td>19,747 (17.5%)</td>
</tr>
<tr>
<td>3</td>
<td>2,957 (19.9%)</td>
<td>469 (19.6%)</td>
<td>12,546 (20.1%)</td>
<td>22,603 (20.0%)</td>
</tr>
<tr>
<td>4</td>
<td>2,551 (17.2%)</td>
<td>407 (16.9%)</td>
<td>11,737 (18.8%)</td>
<td>20,852 (18.5%)</td>
</tr>
<tr>
<td>5 (most affluent)</td>
<td>2,471 (16.6%)</td>
<td>394 (16.4%)</td>
<td>11,991 (19.2%)</td>
<td>20,860 (18.5%)</td>
</tr>
<tr>
<td>Missing</td>
<td>1,021 (6.9%)</td>
<td>157 (6.6%)</td>
<td>3,874 (6.2%)</td>
<td>6,924 (6.1%)</td>
</tr>
<tr>
<td><strong>Unstable housing</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>490 (3.3%)</td>
<td>101 (4.2%)</td>
<td>992 (1.4%)</td>
<td>2,246 (1.8%)</td>
</tr>
<tr>
<td>No</td>
<td>14,358 (96.7%)</td>
<td>2,295 (95.8%)</td>
<td>70,829 (98.6%)</td>
<td>122,251 (98.2%)</td>
</tr>
<tr>
<td><strong>Connection to a primary care provider</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>8,369 (56.4%)</td>
<td>1,446 (60.4%)</td>
<td>28,375 (39.5%)</td>
<td>55,696 (44.7%)</td>
</tr>
<tr>
<td>No</td>
<td>6,479 (43.6%)</td>
<td>950 (39.7%)</td>
<td>43,446 (60.5%)</td>
<td>68,801 (55.3%)</td>
</tr>
<tr>
<td><strong>Primary mental health condition</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anxiety</td>
<td>7,106 (47.9%)</td>
<td>960 (40.1%)</td>
<td>39,520 (55.0%)</td>
<td>66,287 (53.2%)</td>
</tr>
<tr>
<td>Mood</td>
<td>4,904 (33.0%)</td>
<td>947 (39.5%)</td>
<td>19,823 (27.6%)</td>
<td>36,216 (29.1%)</td>
</tr>
<tr>
<td>Schizophrenia</td>
<td>867 (5.8%)</td>
<td>227 (9.5%)</td>
<td>1,023 (1.4%)</td>
<td>3,170 (2.6%)</td>
</tr>
<tr>
<td>Other</td>
<td>1,971 (13.3%)</td>
<td>262 (10.9%)</td>
<td>11,455 (16.0%)</td>
<td>18,824 (15.1%)</td>
</tr>
</tbody>
</table>
### Number of mental health conditions

<table>
<thead>
<tr>
<th></th>
<th>FY2008</th>
<th>FY2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7,152 (48.2%)</td>
<td>837 (34.9%)</td>
</tr>
<tr>
<td>2+</td>
<td>7,696 (51.8%)</td>
<td>1,559 (65.1%)</td>
</tr>
</tbody>
</table>

### Number of comorbid conditions

<table>
<thead>
<tr>
<th></th>
<th>FY2008</th>
<th>FY2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>6,461 (43.5%)</td>
<td>690 (28.8%)</td>
</tr>
<tr>
<td>1</td>
<td>6,165 (41.5%)</td>
<td>1,073 (44.8%)</td>
</tr>
<tr>
<td>2+</td>
<td>2,222 (15.0%)</td>
<td>633 (26.4%)</td>
</tr>
</tbody>
</table>

### Number of hospitalizations (mean/SD)

<table>
<thead>
<tr>
<th></th>
<th>FY2008</th>
<th>FY2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.46 (1.0)</td>
<td>0.55 (1.1)</td>
</tr>
<tr>
<td>SD</td>
<td>0.71 (1.5)</td>
<td>0.84 (1.6)</td>
</tr>
</tbody>
</table>

### Length of stay in hospital (days) (mean/SD)

<table>
<thead>
<tr>
<th></th>
<th>FY2008</th>
<th>FY2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>1.9 (6.6)</td>
<td>2.7 (8.4)</td>
</tr>
<tr>
<td>SD</td>
<td>3.4 (9.9)</td>
<td>4.4 (11.9)</td>
</tr>
</tbody>
</table>

### Emergency department visits (mean/SD)

<table>
<thead>
<tr>
<th></th>
<th>FY2008</th>
<th>FY2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.76 (2.2)</td>
<td>0.88 (2.4)</td>
</tr>
<tr>
<td>SD</td>
<td>1.3 (3.5)</td>
<td>1.4 (3.3)</td>
</tr>
</tbody>
</table>

### Physician visits (mean/SD)

<table>
<thead>
<tr>
<th></th>
<th>FY2008</th>
<th>FY2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>17.6 (17.1)</td>
<td>19.8 (18.4)</td>
</tr>
<tr>
<td>SD</td>
<td>27.5 (22.6)</td>
<td>30.0 (23.3)</td>
</tr>
</tbody>
</table>

### Psychiatrist visits (mean/SD)

<table>
<thead>
<tr>
<th></th>
<th>FY2008</th>
<th>FY2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.86 (4.4)</td>
<td>0.97 (4.9)</td>
</tr>
<tr>
<td>SD</td>
<td>2.4 (8.3)</td>
<td>2.7 (9.3)</td>
</tr>
</tbody>
</table>
Figure 3.2: Classification and regression tree analyses to predict high-cost use in one year, mental health and addiction cohort, Saskatoon, Saskatchewan (n = 124,497; training = 74,622; validation = 49,875)

Validation cohort n=49,875
(87%, 12%)

# physician visits, 1 year prior, <9.8
n=34,406 (not high-cost = 95%; high-cost = 5%)

# physician visits, 1 year prior, >=9.8
n=15,469 (not high-cost = 72%; high-cost=28%)

Health care cost, 1 year prior,
<5119 n=11,405
(80%, 20%)

Health care cost, 1 year prior,
>=5119 n=4,064
(50%, 50%)

# physician visits, 2 years prior, < 16
n=2,081
(64%, 36%)

No comorbidity n=1,016
(73%, 27%)

One or more comorbidities n=1,065
(54%, 46%)

Health care cost, 1 year prior,
<15356 n=833
(80%, 20%)

Health care cost, 1 year prior,
>=15356 n=183
(45%, 55%)

# days in hospital, 1 year prior, <4.1
n=43
(24%, 76%)

# days in hospital, 1 year prior, >=4.1
n=140
(51%, 49%)

# physician visits, 2 years prior, >= 16
n=1983
(35%, 65%)

# physician visits, 2 years prior, >= 30
n=854
(28%, 73%)

No comorbidity n=355
(52%, 48%)

One or more comorbidities, n=774
(35%, 65%)

Health care cost, 1 year prior,
<15356 n=833
(80%, 20%)

Health care cost, 1 year prior,
>=15356 n=183
(45%, 55%)
Figure 3.3: Classification and regression tree analyses to predict persistent high-cost use, mental health and addiction cohort, Saskatoon, Saskatchewan (n = 124,497; training = 74,622; validation = 49,875)

Validation cohort
n=49,875
(not high-cost = 98%, high-cost = 2%)

# physician visits, 1 year prior, < 14.6
n=41,066
(99%, 1%)

# physician visits, 2 years prior, <29.6
n=7219
(94%, 6%)

# physician visits, 1 year prior, < 26.8
n=798
(89%, 11%)

# physician visits, 2 years prior, <67.3
n=947
(78%, 22%)

Primary MH condition not depression
n=689
(81%, 19%)

Primary mental health condition depression
n=258
(69%, 31%)

# physician visits, 1 year prior, <85.4
n=54
(72%, 28%)

# physician visits, 2 years prior, >=67.3
n=93
(69%, 31%)

# physician visits, 1 year prior, >=114.7
n=17
(53%, 47%)

# physician visits, 1 year prior, >=114.7
n=22
(73%, 27%)

# physician visits, 1 year prior, >=14.6
n=8,809
(92%, 8%)

# physician visits, 2 years prior, >=29.6
n=1590
(81%, 19%)

# physician visits, 2 years prior, >=29.6
n=1590
(81%, 19%)

# physician visits, 2 years prior, >=85.4
n=39
(64%, 36%)

# physician visits, 1 year prior, >=114.7
n=22
(73%, 27%)
Table 3.2 Confusion matrix, validation data, predicting one-time high-cost use, mental health and addiction clients, Saskatoon, Saskatchewan (n = 49,875)

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted</th>
<th>Error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>43291</td>
<td>516</td>
</tr>
<tr>
<td>1</td>
<td>4913</td>
<td>115</td>
</tr>
</tbody>
</table>

Table 3.3 Confusion matrix, validation data, predicting persistent high-cost use, mental health and addictions clients, Saskatoon, Saskatchewan (n = 49,875)

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted</th>
<th>Error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>48896</td>
<td>9</td>
</tr>
<tr>
<td>1</td>
<td>962</td>
<td>8</td>
</tr>
</tbody>
</table>
3.5 Discussion

From this study cohort, I developed two predictive models – the first focussed on point-in-time high-cost use and the other on persistent, year after year, high-cost healthcare use. The use of predictive models to flag persons at risk of high-cost use in the future could save healthcare costs (Pingree, 2013; Wrathall & Belnap, 2017).

Creating a model with a high degree of specificity could provide the evidence for public health action to mitigate high-cost use. In contrast to reactive intensive case management strategies, a well-calibrated and validated predictive model could provide evidence for strategies aimed at preventing high-cost use from occurring in the first place.

3.5.1 Summary of findings

The variables most important for prediction were similar between the two outcomes with three notable differences. Predicting one-time high-cost use relied on counts of physician visits, prior health care costs, comorbid conditions, and hospitalizations. Predicting persistent high-cost use also relied on previous physician visit counts, but, with the addition of the primary mental health condition (depression versus other).

The fact that the two predictive models have different variables considered ‘most important’ for prediction is not unexpected. It is reasonable that individuals who are persistently high-cost health care users year after year would be inherently different from episodic high-cost users. It is interesting that in predicting persistent high-cost use the decision tree found a significant difference between primary mental health conditions – potentially demonstrating the chronicity of depression (persistent high-cost use) compared to anxiety (potentially, more episodic).
The findings were similar to previous studies, in particular the comorbid chronic conditions in predicting future high-cost use (Chang et al., 2016; Chechulin et al., 2014; Fleishman & Cohen, 2010; Robst, 2015; Wrathall & Belnap, 2017). Our work builds upon the predictive ability of those models by excluding long-term care residents (known to be high-cost users), using nationally validated case definitions to define comorbid conditions (as opposed to other models reliant on comorbid indices used in hospital), using both historical and recent data, visualizing the predictive algorithm through the use of CART and improving the specificity of overall prediction by predicting future both persistent high-cost use and episodic high-cost use.

Compared to other high-cost healthcare predictive models, our models were comparable in both calibration and specificity (Table 3.2).

It is important to note that good predictors of an outcome are not necessarily causes of the outcome. I make no claim, one way or the other, on the appropriateness of the finding that physician visits increased prior to high-cost use or cause and effect; only that this information could be used to flag persons at-risk of becoming high-cost in the future.

Both models showed excellent specificity; however both models had poor sensitivity. The trade-off between specificity and sensitivity is influenced by the loss function (ie the relative penalty to false positives and false negatives) and the threshold applied to the predicted probability. I could have chosen to modify the sensitivity/specificity balance in the final models, but deliberately did not. If the goal is to accurately predict persistent high-cost use (specificity) – the model functions well. Where it fails, is the large proportion of ‘false negatives’ i.e.) the model misses many individuals who are at-risk. Healthcare systems face the reality of resource constraints. It would be a waste of scare public resources to implement a model into practice with poor specificity, with the potential of erroneously identifying individuals ‘at-risk’, and subsequently
mobilize an action/intervention aimed to prevent their false ‘at-risk’ state. It is reassuring that excellent specificity was achieved. Newer machine learning techniques, namely, the use of random forests and/or tree boosting techniques may have improved model calibration/discrimination (Prasad, Iverson, & Liaw, 2006).

3.5.2 Predictive models in practice

A recent piece in JAMA opined, ‘it seems unlikely that incremental improvements in discriminative performance of the kind typically demonstrated in machine learning research will ultimately drive a major shift in clinical care’ (Shah, Steyerberg, & Kent, 2018). In support of this statement, the authors point to the fact that thousands of models to predict risk of cardiovascular disease are readily available; however, precious few are used in clinical practice. In our view, the interesting qualifier here is, in clinical practice. Perhaps, running predictive models such as these, to predict individuals at risk of future high cost healthcare use, may be more useful at the population level. Perhaps, as the JAMA paper authors suggest, an independent agency would be better positioned to verify calibration/discrimination of predictive models, and, working with health system partners, ensure the predictive analytics deliver on their promise of better value and health outcomes.

The finding that persistent high-cost users’ average number of physician visits increased over time could be a point of intervention. If individuals have evidence of climbing physician visits over time they could be flagged as at risk of persistent high-cost healthcare use. Their physician could even be a partner in the potential intervention point – the data identifying a risk and the primary care provider with the patient-physician relationship using that information for action.
It would be interesting to explore the finding of increased physician visits over time for persistent high cost users in future studies. As a physician visit is one of the least costly interventions in the healthcare system (Canadian Institute for Health Information, 2011) it is unlikely that increased visits is the reason for persistent high-cost use; however, this should be systematically explored. Perhaps the issue is continuity of care— are persistent high-cost users seeing the same primary care provider, or, many different providers from various ‘walk-in’ clinics further compounded by the fact that, despite best efforts of clinical care providers, they are hampered by electronic health information systems that do not ‘talk’ to each other?

In previous work, I found that, taking a myriad of other potentially confounding variables into account, unstable housing was statistically significantly associated with persistent high-cost health care use (Anderson et al., 2018). In that study, I employed logistic regression to explore factors associated with persistent high-cost use. In the current study, I aimed to predict future high-cost use, interestingly, due to the difference in modelling; the variables considered ‘most important’ for prediction in classification and regression trees were not the same variables in the final explanatory models. Either of these modelling techniques could be used for prediction (logistic regression or CART). The decision for employing CART was largely due to the fact that nearly all of the aforementioned predictive models have had good calibration/discrimination, but, evidence to demonstrate they have been used beyond the classical epidemiological sense is limited. By employing CART and its readily interpretable visual outputs, I aimed to create a method more easily implemented into practice by provincial health insurers. Additionally, the variables deemed ‘most important’ for classifying in the decision trees are all commonly available in administrative health data in Canada – adding to the potential the models will be used as data for action.
3.5.3 Interventions for high-cost healthcare use

Among a cohort of mental health and addictions clients, persistent high-cost use can be a flag for system failure. Individuals should not have to consistently use acute care resources, clearly seeking help, but continue to incur high-costs. Upstream, preventative measures aimed to improve the health of the population can help reduce persistent high-cost healthcare use. Intermountain HealthCare in Utah has had success in identifying individuals at risk of future high-cost use and referring them to a Comprehensive Care Clinic. The clinic is aimed at high-cost and those at-risk for high-cost (typically with many chronic conditions and poor continuity of care); the clinic provides increased access to services like primary care, home health, mental health and medication management (Wrathall & Belnap, 2017). In addition, Intermountain has demonstrated quality improvement and cost reduction through their integrated mental health program which seeks to integrate mental healthcare across the care continuum (Reiss-Brennan, Briot, Savitz, Cannon, & Staheli, 2010).

3.5.4 Limitations

This study has several limitations, some inherent to administrative health databases. For example, food insecurity (Tarasuk et al., 2015) has been found to be associated with high-cost use and may have had predictive power in our modelling; however this data was not available for analysis. While I did account for the costs of prescription drugs, I did not examine the types of prescription drugs taken. There may well be a predictive association between drug class and one-time or persistent high-cost use. Community-based mental health services, either publicly funded or private fee-for-service, (individual/group counselling, treatment centres, others) may be associated with persistent high-cost use but not available for analysis. Our study focused on cost;
however, not all healthcare system costs were able to be measured, such as home care, public health, travel costs (air transfers and ground ambulance) and all administrative costs. I found that previous costing data was useful in predicting one-time high-cost use; however, previous healthcare costing data was not available for persistent high-cost users. Inclusion of this costing data may have increased the predictive ability of the model. In defining comorbidities, I relied on available, validated Canadian administrative health database case definitions for select chronic conditions. Lastly, ICD-9 code ‘298’ has been found to be used in systematic error in Saskatchewan. Electronic medical health records in the province erroneously identify code 298 as ‘dementia’ instead of the correct diagnosis of ‘inorganic psychoses’. As dementia is a prevalent condition amongst the elderly there is likely misclassification bias in the MHA category of ‘schizophrenia’ to erroneously include patients with dementia; research into the extent of the error is on-going.

As the current study focused on mental health and addiction patients I did not create methods to define and exclude palliative care patients, ‘alternate level of care patients’ and other groups that may be appropriate high-cost healthcare users – future population-based studies would be recommended to do so. It is recommended future studies focus on clearly delineating ‘actionable’ high-cost use, including the costs savings that could be associated with successful intervention. Future work may also wish to explore machine learning methods that would permit the censoring of deaths, such as hidden Markov models and/or random forest survival models.
Table 3.4 Performance measures, high-cost health care use predictive models

<table>
<thead>
<tr>
<th>Model</th>
<th>Prediction</th>
<th>C-statistic</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chechlin et al (ON, Canada)</td>
<td>Top 5% one year</td>
<td>0.865</td>
<td>57.1</td>
<td>92.5</td>
</tr>
<tr>
<td>Chang et al (CT, US)</td>
<td>Top 5% persistent high cost</td>
<td>0.93</td>
<td>49.7</td>
<td>97.9</td>
</tr>
<tr>
<td>Lu et al (Virginia, US)</td>
<td>Top 10%</td>
<td>0.67</td>
<td>Not available</td>
<td>Not available</td>
</tr>
<tr>
<td>Leininger et al (Survey data, US)</td>
<td>Top 10% (pediatric patients)</td>
<td>0.72</td>
<td>31.1</td>
<td>92.5</td>
</tr>
<tr>
<td>Lauffenburger et al (Aetna insurance enrollees, US)</td>
<td>High cost trajectory group</td>
<td>0.82</td>
<td>Not available</td>
<td>Not available</td>
</tr>
<tr>
<td>Robst et al (FL, US)</td>
<td>Top 1%</td>
<td>Not available</td>
<td>93.1</td>
<td>6.9</td>
</tr>
<tr>
<td>Wrathall et al (Intermountain Health care, US)</td>
<td>Top 15%</td>
<td>0.71</td>
<td>79.0</td>
<td>Not available</td>
</tr>
<tr>
<td>Wrathall et al (Intermountain Health care, US)</td>
<td>Top 15% in 2 of the next 3 years</td>
<td>0.71</td>
<td>48.1</td>
<td>Not available</td>
</tr>
<tr>
<td>Fleishman and Cohen (survey data, US)*</td>
<td>Top 10%</td>
<td>0.854</td>
<td>78.0</td>
<td>Not available</td>
</tr>
<tr>
<td>Current study</td>
<td>Top 10% one year</td>
<td>0.77</td>
<td>23.6</td>
<td>99.0</td>
</tr>
<tr>
<td>Current study</td>
<td>Top 10%, persistent high cost</td>
<td>0.84</td>
<td>11.0</td>
<td>99.0</td>
</tr>
</tbody>
</table>

*best preforming model

3.6 Conclusion

The current study details two well calibrated and specific models developed using classification and regression tree analysis to predict future high-cost use in a cohort of mental health and addiction clients in Saskatoon, Saskatchewan, Canada – both for one time high cost use and persistent high cost use. These models take advantage of healthcare utilization and demographic information readily available in Canadian provincial administrative healthcare databases.

Predictive models can have utility in practice. As discussed above, Intermountain Healthcare uses health data analytics to identify those patients who would benefit most from the intervention and accordingly have stated ‘when I implement evidence-based practices I not only improve clinical outcomes but also do so at a lower cost’ (Pingree, 2013). Similarly, in Saskatchewan, running predictive algorithms on population based healthcare utilization and costing data could provide real, actionable data to prevent high-cost healthcare use in the future.
3.7 References


Waegemakers Schiff, J., Schiff, R., & Schneider, B. (2014). Developing an estimate of supported housing needs for persons with serious mental illnesses [Research article]. https://doi.org/10.1155/2014/245024


3.8 Supplemental material

3.8.1 Cohort and variable creation

All cohort and variable creation for Chapter 3 is identical to Chapter 2, with two exceptions.

Both exceptions occurred due to the learning that occurred as the analyses progressed. In Chapter 2, a total cost per day for each member of the study cohort was calculated, and, population thresholds were calculated as a cost per day. For Chapter 3, however, the data extraction did not permit me to account for individuals who moved in/out of the province and/or died in the total population costing files (used for calculating the population threshold values). Therefore, in Chapter 3, costs were defined per year, not per day. As a result of not being able to account for incomplete follow-up time predictive models were created excluding all deaths. Sensitivity analyses with deaths included and excluded were conducted (see section 3.3).

The second exception was the exclusion of long-term care residents. As the analyses progressed it became clear that LTC residents were nearly entirely persistently high-cost given the cost of a LTC stay; readers will note that, both Chapter 2 and 3 exclude LTC residents from regression modelling, but here in Chapter 3 (unlike in Chapter 2) LTC residents are also excluded from all descriptive analyses.

All other cohort creation and data extraction/manipulation proceeded as outlined in supplemental material 2.7.1 to 2.7.4.
3.8.2 Statistical modeling

In Chapter 3, I was interested in developing a predictive model where the resulting output would be more interpretable to decision-makers. I was also interested in learning a method typically used in ‘big data’ and learning how to partition data into training and validation data sets. To that end, I employed Classification and Regression Tree analyses. The knowledgeable reader will note that the differences between the statistical modeling techniques employed in Chapters 2 and 3 are slight.

Classification and Regression Tree (CART) model are decision trees. Each root node represents a single input variable (x) and a split point on that variable. The leaf nodes of the tree contain an output variable (y) which is used to make a prediction. The resulting output is visual, and, prediction becomes straightforward – a new input transverses over the tree starting at the root node. The CART model effectively makes ‘decisions’ as evident by the splits in the nodes. The optimal split is found by partitioning all of the variables at all possible split points (recursive partitioning). The major components of the CART methodology are: pruning, selection and stopping rules. I conducted many sensitivity analyses using various methods of pruning, selection and stopping – assessing model fit (cross validation misclassification rate) and measures of impurity (Gini score) as the model building progressed.

As a first step, a full tree was built using all of the input data. Further refinements were made through pruning. Cost-complexity methods were used in pruning; due to small cell size/privacy considerations each terminal node was set to include a minimum of 5 individuals. Splitting criteria were based on improvements in entropy, as described by Breiman, using Gini impurity scores in addition to sensitivity/specificity and area under the curve values to evaluate
discriminatory and calibration gains in the final tree (Breiman et al., 1984). A random sampling approach of all study cohort members (see 3.7.1 for details on study cohort creation) was used to partition all modelling data into training (60%) and validation (40%) datasets. Models were fit on training data only; results are reported for validation data only.

Prior to CART analyses, effect selection in the framework of a general linear model with logit link and LASSO selection method (Tibshirani, 1996) was used to limit the number of covariates to those contributing most to outcome measures by choosing the model with the smallest Akaike’s information criterion (AIC) (Akaike, 1973). These variables were then used in classification and regression tree analyses to predict future high-cost health care use (Breiman et al., 1984).

As noted in section 3.7.1, all deaths were excluded from CART modelling.
3.8.3 Co-author affiliations and contributions

Maureen Anderson MSc, PhD (candidate)\textsuperscript{1,2*}, Crawford W Revie MSc, PhD\textsuperscript{1}, Jacqueline M Quail MSc, PhD\textsuperscript{2,7}, Walter Wodchis MA, MAE, PhD\textsuperscript{3,4}, Claire de Oliveira MA, PhD\textsuperscript{3,4,5}, Meric Osman MA\textsuperscript{2}, Henrik Stryhn MSc, PhD\textsuperscript{1}, David L. Buckeridge MD, MSc, PhD\textsuperscript{6}, Nazeem Muhajarine MSc, PhD\textsuperscript{7}, Marilyn Baetz MD\textsuperscript{7,8}, Cordell Neudorf MD, MHSc\textsuperscript{7,9}

\textsuperscript{1} Department of Health Management, University of Prince Edward Island, Charlottetown, Prince Edward Island, Canada
\textsuperscript{2} Health Quality Council (Saskatchewan), Saskatoon, Saskatchewan, Canada
\textsuperscript{3} Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, Canada
\textsuperscript{4} Institute for Clinical Evaluative Sciences, Toronto, Ontario, Canada
\textsuperscript{5} Centre for Addiction and Mental Health, Toronto, Canada
\textsuperscript{6} Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, Quebec, Canada
\textsuperscript{7} Department of Community Health and Epidemiology, College of Medicine, University of Saskatchewan, Saskatoon, Saskatchewan, Canada
\textsuperscript{8} Department of Psychiatry, University of Saskatchewan, Canada
\textsuperscript{9} Saskatchewan Health Authority, Saskatoon, Saskatchewan, Canada

\textbf{MA} – Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Software, Validation, Visualization, Writing: original draft preparation, editing, Final approval and Agreement to be accountable

\textbf{CR} – Conceptualization, Methodology, Visualization, Supervision, Writing: review and editing, Final approval and Agreement to be accountable

\textbf{JQ} – Conceptualization, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Writing – review and editing, Final approval and Agreement to be accountable

\textbf{WW} – Conceptualization, Funding Acquisition, Methodology, Supervision, Writing: review, Final approval and Agreement to be accountable

\textbf{CO} – Conceptualization, Funding Acquisition, Methodology, Supervision, Writing – review and editing, Final approval and Agreement to be accountable

\textbf{MO} – Data Curation, Formal Analysis, Investigation, Methodology, Software, Validation, Writing: review and editing, Final approval and Agreement to be accountable

\textbf{HS} – Conceptualization, Methodology, Supervision, Validation, Writing – review and editing, Final approval and Agreement to be accountable

\textbf{DB} – Conceptualization, Validation, Supervision, Final approval and Agreement to be accountable
NM – Funding Acquisition, Writing – review and editing, Final approval and Agreement to be accountable

CN – Conceptualization, Funding Acquisition, Resources, Supervision, Writing: review and editing, Final approval and Agreement to be accountable
4. DEFINING ‘ACTIONABLE’ HIGH-COST HEALTH CARE USE: RESULTS USING THE CANADIAN INSTITUTE FOR HEALTH INFORMATION POPULATION GROUPING METHODOLOGY

Maureen Anderson MSc, PhD1,2*, Crawford W Revie MSc, PhD1,3, Henrik Stryhn PhD1, Cordell Neudorf MD, MHSc4,5, Yvonne Rosehart BA6, Wenbin Li MSc, PhD7,7, Meriç Osman MA2, David L Buckeridge MD, MSc, PhD8, Laura C Rosella MSc, PhD7,9-11, Walter P Wodchis MA, MAE, PhD7,10,11


4.1 Abstract

Background: A small proportion of the population consumes the majority of health care resources. High-cost health care users are a heterogeneous group. I aim to segment a provincial population into relevant homogenous sub-groups to provide actionable information on risk factors associated with high-cost health care use within sub-populations.

Methods: The Canadian Institute for Health Information (CIHI) Population Grouping methodology was used to define mutually exclusive and clinically relevant health profile sub-groups. High-cost users (>=90th percentile of health care spending) were defined within each sub-group. Demographic, socio-economic status, health status and health care utilization variables were explored for potential association with high-cost use. Following univariate and bivariate analyses, multivariable logistic regression models were constructed for the costliest health profile groups.

Results: From 2015-2017 a total of 1,175,147 individuals were identified for study. High-cost users consumed 41% of total health care resources. Average annual health care spending for individuals not high-cost were $642; high-cost users were $16,316. The costliest health profile groups were ‘long-term care’, ‘palliative’, ‘major acute’, ‘major chronic’, ‘major cancer’, ‘major newborn’, ‘major mental health’ and ‘moderate chronic’. These groups were modelled to better understand risk factors associated with high-cost use. Both ‘major acute’ and ‘major cancer’ health profile groups were largely explained by measures of health care utilization and multimorbidity. In the remaining costliest health profile groups modelled, ‘major chronic’, ‘moderate chronic’, ‘major newborn’ and ‘other mental health’, a measure of socio-economic status, low neighbourhood income, was statistically significantly associated with high-cost use.

Interpretation: Model results point to specific, actionable information within clinically meaningful subgroups to reduce high-cost health care use. Population segmentation methods, and more specifically, the CIHI Population Grouping Methodology, provide specificity to high-cost health care use; informing interventions aimed at reducing health care costs and improving population health.
4.2 Introduction

Health care systems in Canada and around the world are challenged with increased costs. Health care expenditure in nearly every developed country meets or exceeds 10% of gross domestic product (World Health Organization, 2015). Evidence has long demonstrated that a small proportion of the population (<10%) accounts for the majority (50-70%) of total health care spending (Densen et al., 1959; Roos et al., 2003; Roos et al., 1989; Stuart & Weinrich, 1998); individuals commonly referred to as ‘high-cost users’.

High-cost users are a heterogeneous population. From 1985 (n = 79) to 2018 (n = 1,198) the number of PubMed citations involving the term ‘high-cost use’ has increased over 15 times (PubMed - NCBI, 2018). In general, previous studies have found high-cost health care use to be associated with: 1) complex, multiple chronic conditions (multi-morbidity), 2) catastrophic illness (for example, motor vehicle accident or major cancer), 3) high costs at end-of-life, 4) mental health and addictions, 5) institutional living, and 6) various indicators of lower socio-economic status (for example, food insecurity and poverty) – and combinations thereof. A recent systematic review on high-cost health care users identified similar patterns: multi-morbidity, mental health and addictions, increasing age, end-of-life care and socio-economic status were the predominant factors associated with high-cost use across 55 countries globally (Wammes et al., 2018).

One of the pioneers in understanding high-cost health care use is Dr. Jeffrey Brenner and his team in Camden, New Jersey. They found high-cost users were predominately located in two high-rise urban apartment buildings in an impoverished neighbourhood in Camden City. Case management strategies, such as providing ‘wrap-around’ care for complex patients, aimed
specifically at the identified population both improved health outcomes and reduced cost (Brenner, 2013; Kaufman et al., 2014).

When public health action is focused on a relatively small population, such as the one in Camden, New Jersey (Brenner, 2013; Kaufman et al., 2014) it is easier for the insurers to describe and create policies aimed at improving health outcomes and reducing inappropriate health care costs. For example, an interdisciplinary team of care providers in Camden were able to co-ordinate their services within a specific geographical location and a relatively homogenous population (low socio-economic status). However, when a population is more diffuse and broad, as is the case with publicly insured provincial populations in Canada, describing and creating policies aimed at reducing high-cost health care use are more difficult.

To that end I aimed to understand, by specific mutually exclusive health profile groups, factors driving high-cost health care use within each health profile group under study. By defining ‘high-cost use’ within sub-groups, better specificity to the high-cost definition is achieved. For example, if one were to define ‘high-cost use’ as those in the >=90th percentile of spending in the provincial population overall, factors that are related to high-cost use in general would be understood; however, nuances of factors that may be associated with high-cost use within a specific subset, such as acute disease versus chronic disease, would be lost.

In the current study, ‘actionable’ is defined as quantitatively understanding risk factors associated with high-cost use, within sub-populations, including variables that are typically included in modelling high-cost use but less amenable to change (for example, age) with variables that may be more amenable to change and therefore ‘actionable’ through targeted programmatic or policy interventions (for example, socio-economic status).
Achieving equity in health is one of the key drivers of this research; high-cost health care users are more costly to the health care system, but, is this higher cost a flag indicating where the system could improve? Is high-cost indicative of a health care system that is inequitable? Taking potentially confounding factors into account, who is more at risk of high-cost use?

In order to achieve study objectives, I used a population segmentation method recently developed by the Canadian Institute for Health Information (CIHI). The population grouping methodology (‘Pop Grouper’) builds clinical and demographic profiles for each person in a population, including health system non-users.

CIHI’s population grouping methodology enables health system planners and policy-makers to use evidence to support decision-making, help CIHI and its clients monitor population health and diseases, predict health care utilization patterns and explain variations in health care resource use, provide a foundation for funding models, and, facilitate standardization of populations for among province analyses (Hatcher, Richards, Homan, & Zhu, 2017) (see Methods and Supplemental Material for details).

Given rising health care costs, increased demand from growing elderly populations with multi-morbidity, and, health consuming more and more of the total gross domestic product in many countries around the world, policy makers and health researchers have, reasonably, been looking to the population of ‘high-cost users’ for cost savings.

4.2.1 Theoretical framework
As this research is based on health care utilization data, it is important to understand the theoretical framework surrounding health care utilization in general. The ‘Behavioural Model of Health Services Use’ developed by RM Andersen in 1968 (updated in 1995 and renamed the
‘Andersen Health care Utilization Model’) is considered foundational work in this area (Andersen, 1995).

This conceptual model describes the factors that lead to the use of health services. According to the model, there are different dynamics that affect an individual’s health care utilization. Andersen categorizes predictors into three categories: need, enabling and predisposing factors (Andersen, 1995; Babitsch, Gohl, & von Lengerke, 2012). Briefly, predisposing characteristics are those that predispose individuals to use or not use health care services (such as age, sex, ethnicity); enabling characteristics are those that either increase or decrease the likelihood of health care service use (such as trust of health care system, income level) and need characteristics are those related to health care service, both perceived and actual need, such as the presence of chronic conditions.

The model makes a distinction between equitable and non-equitable access to health care services. Equitable access is driven by predisposing factors and need. Inequitable access is driven by predisposing and enabling factors. For example, an individual who believes Western medical health care services are beneficial to their perceived need are more likely to seek care; however, the ability to access services might vary based on ethnicity, sexual orientation, economic status, and other factors. Need, predisposing and enabling factors, which can change over time, can all affect an individuals’ use of health care services, including regression to the mean.

**Study objectives:** 1) To describe and define high-cost health care use in the provincial population of Saskatchewan, Canada and, 2) From the perspective of a provincial government
funding health care in Canada, to explore risk factors associated with high-cost health care use within specific health profile groups.

To the best of our knowledge, this is the first study to segment the population into clinically meaningful sub-groups, define high-cost use within each sub-group and include a measure of socio-economic status in multivariable regression models.

4.3 Methods

Due to the availability of administrative health data to the study through in-kind support from host institutions and data-sharing agreements, this study focuses on the provincial population of Saskatchewan, Canada (population ~1.2 million); a Canadian province with a central provincial health insurer. Most Saskatchewan residents receive provincial health care benefits -- with the exception of less than 1% of the population for whom benefits are provided by the federal government (members of the armed forces and federal penitentiary inmates) (Downey et al., 2006). Every provincial resident with a valid health service number (‘health card’) for at least one day from April 1 2015 to March 31 2017 was eligible for study.

**Population segmentation method:** The Canadian Institute for Health Information (CIHI) population grouping methodology was used to segment the study population into clinically meaningful and mutually exclusive health profile groups (Hatcher et al., 2017) (Supplemental material). Diagnosis-related grouping (DRG) methodologies are used throughout the world. American DRGs are readily available for use in Canada; however, these models have two significant drawbacks: 1) they were not designed using Canadian data and 2) they are intellectual property of the company that produce them, and, as such, are relatively expensive to purchase.
In brief, CIHI’s population grouping methodology uses a combination of administrative health databases and provincial health registry systems to ‘tag’ each resident with any of the 239 health conditions. These binary tags (0/1) form the ‘building block’ of the grouping methodology and are not mutually exclusive; an individual can have any number of applicable health conditions. The presence of health conditions are determined by linking data from hospitalizations, physician visits, hospital day surgeries, emergency department visits and long-term care. The method uses 24 months of data to determine health conditions (n = 239), branches (n = 164) and health profile groups (n = 16). Please see ‘Supplemental material’ for addition details.

Unlike common comorbidity measures such as the Charlson or Elixhauser indexes (Southern, 2004) the Pop Grouper does not require an individual to have been hospitalized; Case Mix Groups (a DRG) similarly require the individual to have been hospitalized. Pop Grouper categories are available for every resident, including non-users of the health care system. Unlike proprietary DRGs or other population segmentation methods, Pop Grouper was developed in Canada, underwent extensive validation exercises using Canadian data and its development made use of Canadian clinical content experts (Hatcher et al., 2017).

During the course of the study, I identified a further ‘health profile group’, ‘Long-term care (LTC) resident’; these study subjects were identified through the use of the RAI-LTC 2.0 (see ‘Databases’ for details). After health profile grouping, logistic models were constructed for each profile group to understand factors associated with ‘high-cost health care use’ (defined as \( \geq 90^{th} \) percentile within health profile groups). Independent variables in the models were those commonly available in administrative health databases (demographic and health care utilization data). Measures of equity were defined using the Institute National Santé Publique du Quebec’s
area-based deprivation index; this method divides neighbourhoods into quintiles according to levels of total deprivation (Pampalon and Gamache, 2011).

I made use of Andersen’s behavioural model and theoretical framework for health services use in describing predictors of interest. Andersen categorizes predictors into three categories: need, enabling and predisposing factors (Andersen, 1995; Babitsch et al., 2012). Briefly, *predisposing* characteristics are those that predispose individuals to use or not use health care services (such as age, sex, ethnicity); *enabling* characteristics are those that either increase or decrease the likelihood of health care service use (such as trust of health care system, income level) and *need* characteristics are those related to health care service need, such as the presence of chronic conditions. Where the data was collected in routine administrative health databases these variables were accounted for in the models.

**Databases:** Detailed descriptions of Saskatchewan Health databases are available elsewhere (Downey et al., 2006). In brief, demographic characteristics, location of residence, and neighbourhood income quintile were extracted from the *Personal Health Registration System (PHRS)* for individuals with >= 1 day of valid health insurance coverage within the study period. Hospital data were extracted from the *CIHI-Discharge Abstract Database (DAD)* and includes inpatient and day surgery records for the province of Saskatchewan. Out-of-province hospitalizations for Saskatchewan residents were included; transfers were included but ‘counted’ as one hospitalization. The International Classification of Diseases (ICD), 10th revision, Canadian Version (ICD-10-CA) was used in the DAD to record up to 25 diagnoses at discharge, including the primary responsible diagnosis for that hospitalization. Data on physician services are contained in the *Medical Services Claims Database*. Physicians paid on a fee-for-service
basis submit billing claims to the provincial health ministry; a single diagnosis using a three-digit ICD-9 code is recorded on each claim. The same single diagnosis is reported on every claim submitted for a single visit – multiple service claims by the same physician, same patient, and same day were counted as one ‘visit’. Salaried physicians can submit billing claims for administrative purposes (shadow billing); however, claims from salaried physicians involve under-reporting resulting in fewer physician claims in the data than in practice (Saskatchewan Ministry of Health, 2017). Residents of long-term care facilities were defined according to the provincial Resident Assessment Instrument-Minimum Dataset for long-term care facilities (RAI-LTC 2.0). Home care data was accessed through the Resident Assessment Instrument-Minimum Dataset for home care facilities (RAI-HC). Emergency department (ED) visit data was recorded in National Ambulatory Care Reporting System (NACRS); however, for the study period <50% of the provincial emergency departments were NACRS-reporting facilities resulting in under-reporting of ED visits. Prescription drug data was used to calculate total government paid prescription drug costs at the individual-level (see ‘Outcome’), but, drug data was not extracted as health care utilization.

Death data were obtained using a ‘derived death file’ (combination of death data from various administrative health databases) created by the Saskatchewan Health Quality Council for research purposes. Data were stripped of nominal variables (i.e.) name and health services number, and, linked at the individual level using a unique non-identifiable number generated by eHealth Saskatchewan.

**Outcome:** Total health care costs were calculated for each individual for the study period. Health care costs were a sum, at the individual-level, of total costs associated with
hospitalizations (both in-patient and day surgery), physician visits, emergency department visits, long-term care beds and prescription drugs. Hospitalization costs were calculated by multiplying the CIHI Resource Intensity Weight (RIW) value at the patient-level by the CIHI derived value ‘cost of a standard hospital stay’. In the absence of RIWs for LTC in Saskatchewan, long-term care costs were calculated using a per-diem value approach (per-diem value calculated using LTC actual expenditure, resident fees and number of LTC beds). Total government-paid costs were used in total health care costing for each of the other databases. Costing variables were assigned to their fiscal year of occurrence. Total health care costs attributed to each individual in the study population were calculated using the person-level costing methodology developed by Wodchis et al. Briefly, this methodology provides guidance on how to identify unit costs associated with individual health care utilization of emergency departments, hospitalizations, physician visits, long-term care and prescription drugs. The method includes the ability to combine these costs with utilization data from administrative health databases; providing a measure of direct health care costs incurred by government. Each of the administrative health databases calculates cost in a different way – I then unified the calculation across the databases by combining them as the ‘total cost per person’.

In the current study, I defined high-cost status within each health profile group. Therefore, the value above which a person would be considered ‘high-cost’ would be different between the health profile groups; however, this value was consistently defined as the 90th percentile of total cost within the health profile group. Although I would have preferred to account for death/relocation during the study period as individuals with less follow-up time would have lower costs, coverage days were not readily available for analysis (see Limitations). For each health profile group (n = 16) ‘high-cost users’ were uniquely identified within the health profile
group as those individuals exceeding the 90th percentile of group-specific total health care costs (n = 10% within each health profile group). The health profile groups with the highest total costs (n = 6, excluding palliative care and long-term care residents) were modelled to assess factors associated with high-cost use: ‘Major acute’, ‘Major chronic’, ‘Major cancer’, ‘Moderate chronic’, ‘Major Newborn’ and ‘Other Mental Health’.

Predictors: Health care utilization variables are, in most instances, directly related to health care costs. Comorbid conditions, age, sex, geographical location and socio-economic status variables are, in most instances, potentially confounding variables in the relationship between utilization and cost (Wammes et al, 2018).

Demographic and socio-economic variables (geographic location, age, sex; predisposing factors; neighbourhood income quintile; predisposing and enabling factors) were defined as of study index date (April 1 2015) and extracted from the provincial Personal Health Registry System. Chronic comorbid conditions were identified using the health condition ‘tags’ embedded within the Pop Grouper (need factors). Health care utilization variables - number of physician visits, number of emergency department visits, number of hospitalizations, home care visits, long-term care residency, length of stay in hospital and alternate level of care hospital days - were extracted from the relevant administrative health database, by fiscal year, for the duration of the entire study period (April 1 2015 to March 31 2017). ‘Alternate level of care’ hospital days refer to those days spent in hospital where the level of acuity is not needed for the patient, however, there does not exist a suitable place for the patient to be transferred to; most typically, individuals are awaiting a LTC bed (Nord, 2009).
Health profile groups were defined using the CIHI Population Grouping Methodology software. A detailed description of this methodology precedes this section. Cancer, dialysis, mental health and neonatal intensive care conditions were defined using each relevant binary health condition ‘tags’. Individuals ‘tagged’ with $\geq 1$ of these health conditions were considered to have the relevant condition(s). Multiple chronic conditions were defined using these same binary health conditions tags for minor, moderate and major chronic disease health profile groups. Based on descriptive analyses of the data, any individual with $\geq 3$ of these chronic condition ‘tags’ were considered to have multiple chronic conditions. Categorical variables of health care utilization (‘high’ number of visits versus not) were defined as any individual $\geq 75^{th}$ percentile of total study population health care utilization (this equated to: physician visits $\geq 23$; emergency department visits $\geq 1$ and hospitalizations $\geq 1$ over the study period). The $75^{th}$ percentile cut-off was used as descriptive statistics indicated the continuous count variables sharply rose at this value; thus marking a potential difference between ‘high’ and ‘low’ (subsequently modelled to understand the effect). See Table 1 for descriptive statistics of all study variables.

**Statistical Analyses:** Following univariate and bivariate analyses, multivariable regression modelling was used to delineate factors associated with high-cost use; as defined within each health profile group. Logistic regression model effect selection was achieved by limiting the number of covariates to those contributing most to outcome measures by choosing the model with the smallest Akaike’s information criterion (AIC) (Akaike, 1973). All interaction terms were explored analytically; only those deemed biologically plausible by clinician contributors and previous studies were included in the models. Where missing values were present in the PHRS (income quintile and location of residence) a categorical value of ‘missing’ was created in order to allow for sensitivity analyses, with and without the subjects with missing data.
Sensitivity analysis, with and without, missing data were conducted. All analyses were conducted using SAS© Enterprise Guide version 7.1 (SAS Institute Inc., 2017).

The study proposal underwent ethical review and approval by the University of Saskatchewan Biomedical Research Ethics Board and the University of Prince Edward Island Research Ethics Board for research involving human subjects.
4.4 Results

A total of 1,175,147 individuals, residents of Saskatchewan, excluding residents of long-term care, with health insurance coverage of at least one day from April 1 2015 to March 31 2017 and person-level costing data were identified for study (Figure 1). Regardless of health condition(s) present LTC residents were consistently high-cost health care users and therefore excluded from further study. In descriptive analyses of the provincial population, high-cost health care users (n = 117,512) were more likely to be older (80+ years), female, residents of rural Saskatchewan, lower income, have more than one chronic condition and die during the study period compared to non-high cost users. Compared to non-high cost users, high-cost users were more likely to have higher health care utilization, such as, home care services, be hospitalized (with longer lengths of stay and more ‘alternate level of care’ hospitalizations), visit the emergency department and have increased physician visits (Table 1); these findings varied by each health profile group (data not shown).

In terms of health profile group categories, there were differences between high-cost users within health profiles groups compared to general population high-cost users. The majority of the general study population, regardless of cost, was grouped into ‘minor acute disease’, ‘health system non-user’, ‘minor chronic disease’ and ‘moderate chronic disease’. High-cost users in the study population were mainly grouped into ‘moderate chronic’, ‘major chronic disease’, ‘major acute disease’, ‘moderate acute’, and ‘obstetrics’ (71.9%). Overall, high-cost users in Saskatchewan comprised 10% of the study population but accounted for 41% of total health care costs in the study period.

Factors associated with high-cost health care use are detailed by health profile group.
**Major chronic disease health profile group**

Individuals categorized as ‘major chronic disease’ over the study period were assessed for risk factors associated with their high-cost use. Using available variables and taking known risk factors for high-cost use into account, low income was significantly associated with high-cost health care use (OR=29.4; 95% CI: 19.4-44.7); low income interacted with length of hospital stay exceeding three days in the study period. Individuals who did not have low income but did have a length of stay >=3 days had a lower risk of high-cost health care use compared to individuals with low neighbourhood income and a length of hospital stay >=3 days (OR: 15.3; 95% CI: 12.9-18.1).

**Moderate chronic disease health profile group**

Under the ‘moderate chronic disease’ health profile group, those with low neighbourhood income and high numbers of hospitalizations were more likely to be high-cost users, compared to low numbers of hospitalizations and not low income (OR=4.8; 95% CI: 4.3-5.2).

**Major acute disease health profile group**

In the major acute disease health profile group, high-cost use was associated with long hospital stays (OR: 35.5; 95% CI: 26.8-47.1). In addition, among those with multiple chronic conditions and high numbers of physician visits had a reduced risk of high-cost health care use (OR: 2.1 versus OR: 3.8; 95% CI: 2.3-6.2).
**Major cancer disease health profile group**

In the major cancer health profile group, high-cost use was associated with health care utilization variables (emergency department, home care hospitalizations and length of stay) and the presence of multiple chronic conditions.

**Major newborn health profile group**

In the ‘major newborn’ health profile group, newborn babies in low-income neighbourhoods were at increased risk of high-cost use – having increased visits to a physician reduced this risk (OR=1.44 versus OR=0.81).

**Other mental health profile group**

I aimed to model the costliest health profile groups, in terms of average cost, which would include the ‘major mental health’ group. However, in Saskatchewan, a systematic error in electronic medical databases throughout the province incorrectly defines ICD-9 code 298 as ‘dementia’ as opposed to the correct definition of ‘inorganic psychoses’. Due to this error, the ‘major mental health’ health profile group has a large preponderance of individuals >=80 years with high health care utilization; likely indicative of dementia patients. As mental health and addiction is a known driver of high-cost health care use (Wammes et al., 2018), I include results for the ‘other mental health’ health profile group, not one of the costliest groups, but, a representation of a mental health and addictions health profile group.

In the ‘other mental health’ category low income interacted with high physician visits; those with low income and high physician visits had an increased risk of high-cost use (OR=5.2; 95% CI:
4.8-5.5) compared to individuals not low income with high physician visits (OR = 3.35; 95% CI: 3.02-3.72).
Figure 4.1: Study cohort inclusion/exclusion criteria

- Resident of Saskatchewan with >= 1 day health care coverage between April 1 2015 - March 31 2017 ($3.8 billion)  
  \( n = 1,240,844 \)

- Person-level cost data missing  
  \( n = 1,185,896 \)

- Exclude long-term care residents ($248 million)  
  \( n = 1,175,147 \)

- Final study cohort (~$3.5 billion)  
  \( n = 1,175,147 \)
Table 4.1 Descriptive epidemiology, demographics and health care utilization variables, by cost category, study population, April 1 2015 – March 31 2017 (n = 1,175,147)

<table>
<thead>
<tr>
<th></th>
<th>Lowest 90%, Saskatchewan (n = 1,057,635)</th>
<th>Top 10%, Saskatchewan (n = 117,512)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-79 years</td>
<td>1,022,482 (96.7%)</td>
<td>100,667 (85.7%)</td>
</tr>
<tr>
<td>80+ years</td>
<td>35,153 (3.3%)</td>
<td>16,845 (14.3%)</td>
</tr>
<tr>
<td><strong>Sex</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>542,573 (51.3%)</td>
<td>49,906 (42.5%)</td>
</tr>
<tr>
<td>Female</td>
<td>515,062 (48.7%)</td>
<td>67,606 (57.5%)</td>
</tr>
<tr>
<td><strong>Geographic location</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Urban</td>
<td>711,058 (67.2%)</td>
<td>77,693 (66.1%)</td>
</tr>
<tr>
<td>Rural</td>
<td>317,614 (30.0%)</td>
<td>36,656 (31.2%)</td>
</tr>
<tr>
<td>Missing</td>
<td>28,963 (2.7%)</td>
<td>3,163 (2.7%)</td>
</tr>
<tr>
<td><strong>Neighbourhood income</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 (least affluent)</td>
<td>225,650 (21.3%)</td>
<td>26,426 (22.5%)</td>
</tr>
<tr>
<td>2</td>
<td>192,838 (18.2%)</td>
<td>21,918 (18.7%)</td>
</tr>
<tr>
<td>3</td>
<td>177,797 (16.8%)</td>
<td>19,685 (16.8%)</td>
</tr>
<tr>
<td>4</td>
<td>202,522 (19.2%)</td>
<td>22,363 (19.0%)</td>
</tr>
<tr>
<td>5 (most affluent)</td>
<td>176,755 (16.7%)</td>
<td>18,710 (15.9%)</td>
</tr>
<tr>
<td>Missing</td>
<td>82,073 (7.8%)</td>
<td>8,410 (7.2%)</td>
</tr>
<tr>
<td><strong>Health profile category</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Palliative</td>
<td>1,456 (0.1%)</td>
<td>4,754 (4.1%)</td>
</tr>
<tr>
<td>Major acute</td>
<td>19,010 (1.8%)</td>
<td>15,723 (13.4%)</td>
</tr>
<tr>
<td>Major chronic</td>
<td>20,292 (2.0%)</td>
<td>17,511 (14.9%)</td>
</tr>
<tr>
<td>Major newborn</td>
<td>1,519 (0.1%)</td>
<td>1,159 (1.0%)</td>
</tr>
<tr>
<td>Major mental health</td>
<td>11,031 (1.0%)</td>
<td>6,581 (5.6%)</td>
</tr>
<tr>
<td>Major cancer</td>
<td>4,803 (0.5%)</td>
<td>3,908 (3.3%)</td>
</tr>
<tr>
<td>Moderate acute</td>
<td>63,528 (6.0%)</td>
<td>13,492 (11.5%)</td>
</tr>
<tr>
<td>Moderate chronic</td>
<td>87,826 (8.3%)</td>
<td>24,950 (21.2%)</td>
</tr>
<tr>
<td>Other cancer</td>
<td>4,595 (0.4%)</td>
<td>1,321 (1.1%)</td>
</tr>
<tr>
<td>Other mental health</td>
<td>58,587 (5.5%)</td>
<td>3,794 (3.2%)</td>
</tr>
<tr>
<td>Obstetrics</td>
<td>22,399 (2.1%)</td>
<td>12,794 (10.9%)</td>
</tr>
<tr>
<td>Minor acute</td>
<td>424,832 (40.2%)</td>
<td>7,087 (6.0%)</td>
</tr>
<tr>
<td>Minor chronic</td>
<td>146,122 (13.8%)</td>
<td>3,731 (3.2%)</td>
</tr>
<tr>
<td>Healthy newborn</td>
<td>9,601 (0.9%)</td>
<td>412 (0.4%)</td>
</tr>
<tr>
<td>Health system user, no health conditions</td>
<td>52,527 (5.0%)</td>
<td>83 (0.1%)</td>
</tr>
<tr>
<td>Health system non-user</td>
<td>129,507 (12.2%)</td>
<td>212 (0.2%)</td>
</tr>
<tr>
<td><strong>Multi-morbidity (&gt;=3 conditions)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>96,570 (9.1%)</td>
<td>54,392 (46.3%)</td>
</tr>
<tr>
<td>No</td>
<td>961,065 (90.9%)</td>
<td>63,120 (53.7%)</td>
</tr>
<tr>
<td><strong>Died during study period</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>9,244 (0.9%)</td>
<td>11,686 (9.9%)</td>
</tr>
<tr>
<td>No</td>
<td>1,048,391 (99.1%)</td>
<td>105,826 (90.1%)</td>
</tr>
<tr>
<td>Health Indicator</td>
<td>Lowest 90%, Saskatchewan (n = 1,057,635)</td>
<td>Top 10%, Saskatchewan (n = 117,512)</td>
</tr>
<tr>
<td>---------------------------------------------</td>
<td>------------------------------------------</td>
<td>-------------------------------------</td>
</tr>
<tr>
<td>Home care client</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>47,228 (4.5%)</td>
<td>34,357 (29.2%)</td>
</tr>
<tr>
<td>No</td>
<td>1,010,407 (95.5%)</td>
<td>83,155 (70.8%)</td>
</tr>
<tr>
<td>Number of hospitalizations (mean/SD)</td>
<td>0.22 (0.6)</td>
<td>2.1 (2.1)</td>
</tr>
<tr>
<td>Length of stay (LOS) in hospital (days)</td>
<td>0.72 (7.1)</td>
<td>13.0 (33.1)</td>
</tr>
<tr>
<td>Alternate level of care</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>2,521 (0.2%)</td>
<td>4,340 (3.7%)</td>
</tr>
<tr>
<td>No</td>
<td>1,055,114 (99.8%)</td>
<td>113,172 (96.3%)</td>
</tr>
<tr>
<td>Mental health condition</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>108,044 (10.2%)</td>
<td>29,795 (25.4%)</td>
</tr>
<tr>
<td>No</td>
<td>949,591 (89.8%)</td>
<td>87,717 (74.7%)</td>
</tr>
<tr>
<td>Dialysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>1,657 (0.2%)</td>
<td>3,287 (2.8%)</td>
</tr>
<tr>
<td>No</td>
<td>1,055,978 (99.8%)</td>
<td>114,225 (97.2%)</td>
</tr>
<tr>
<td>Emergency department visits (mean/SD)</td>
<td>0.2 (0.6)</td>
<td>0.8 (2.4)</td>
</tr>
<tr>
<td>Family physician visits (mean/SD)</td>
<td>4.4 (5.6)</td>
<td>12.8 (12.4)</td>
</tr>
<tr>
<td>Specialist physician visits (mean/SD)</td>
<td>2.7 (6.1)</td>
<td>16.0 (23.2)</td>
</tr>
<tr>
<td>Total physician visits &gt;=23/yr</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>206,751 (19.6%)</td>
<td>96,693 (82.3%)</td>
</tr>
<tr>
<td>No</td>
<td>850,884 (80.5%)</td>
<td>20,819 (17.7%)</td>
</tr>
<tr>
<td>History of hospitalizations &gt;=1/yr</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>41,575 (3.9%)</td>
<td>59,502 (50.6%)</td>
</tr>
<tr>
<td>No</td>
<td>1,016,060 (96.1%)</td>
<td>58,010 (49.4%)</td>
</tr>
<tr>
<td>Average annual health care cost ($) (mean/SD)</td>
<td>$642 ($895)</td>
<td>$16,316 ($23,992)</td>
</tr>
<tr>
<td>Total health care cost ($) (%)</td>
<td>$2,049,772,060 (58.7%)</td>
<td>$1,442,545,027 (41.3%)</td>
</tr>
</tbody>
</table>
Table 4.2: Mean health care costs, by health profile category and high-cost use, Saskatchewan, excluding long-term care residents, April 1 2015 to March 31 2017 (n = 1,175,147)

<table>
<thead>
<tr>
<th>Health profile category*</th>
<th>n</th>
<th>Cost per person (mean, SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Palliative</td>
<td>6,210</td>
<td>$30,301 (39,692)</td>
</tr>
<tr>
<td>Major Newborn</td>
<td>2,678</td>
<td>$14,714 (39,401)</td>
</tr>
<tr>
<td>Major Acute</td>
<td>34,733</td>
<td>$12,062 (26,737)</td>
</tr>
<tr>
<td>Major Chronic</td>
<td>37,803</td>
<td>$11,663 (23,434)</td>
</tr>
<tr>
<td>Major Cancer</td>
<td>8,711</td>
<td>$9,652 (15,886)</td>
</tr>
<tr>
<td>Major Mental Health</td>
<td>17,612</td>
<td>$8,389 (17,655)</td>
</tr>
<tr>
<td>Moderate Chronic</td>
<td>112,776</td>
<td>$3,788 (7,064)</td>
</tr>
<tr>
<td>Other Cancer</td>
<td>5,916</td>
<td>$3,782 (6,320)</td>
</tr>
<tr>
<td>Obstetrics</td>
<td>35,193</td>
<td>$3,489 (4,273)</td>
</tr>
<tr>
<td>Moderate Acute</td>
<td>77,020</td>
<td>$2,843 (4,820)</td>
</tr>
<tr>
<td>Healthy Newborn</td>
<td>10,013</td>
<td>$2,023 (1,669)</td>
</tr>
<tr>
<td>Other Mental Health</td>
<td>62,381</td>
<td>$1,442 (3,306)</td>
</tr>
<tr>
<td>Minor Chronic</td>
<td>149,853</td>
<td>$918 (2,188)</td>
</tr>
<tr>
<td>Minor Acute</td>
<td>431,919</td>
<td>$518 (1,731)</td>
</tr>
<tr>
<td>Health System User, no  health conditions</td>
<td>52,610</td>
<td>$160 (1,221)</td>
</tr>
<tr>
<td>Health System non-user</td>
<td>129,719</td>
<td>- (-)</td>
</tr>
</tbody>
</table>

*Mutually exclusive health profile categories assigned by highest resource intensity April 1 2015 to March 31 2017
### Tables 3-8: Logistic regression models comparing high-cost health care users and not high-cost use, excluding long-term care residents, by health profile category, Saskatchewan, April 1 2015 to March 31 2017 (n = 1,175,147)

*Note:* the following tables contain the OR comparisons considered of primary interest from the models.

Table 4.3 High-cost use within major chronic health profile group, odds ratios (Total n = 37,803; High-cost users n = 3,781)

<table>
<thead>
<tr>
<th>Focus setting</th>
<th>Comparison setting</th>
<th>OR</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Home care (=yes)</td>
<td>Home care (=no)</td>
<td>1.96</td>
<td>1.82-2.12</td>
</tr>
<tr>
<td>LOS = y; low income = y</td>
<td>LOS = n; low income = y</td>
<td>29.42</td>
<td>19.37-44.69</td>
</tr>
<tr>
<td>LOS = y; low income = n</td>
<td>LOS = n; low income = n</td>
<td>15.25</td>
<td>12.86-18.09</td>
</tr>
<tr>
<td>Multi chronic dx = y; High Dr. visits = y</td>
<td>Multi chronic dx = y; High Dr. visits = n</td>
<td>4.99</td>
<td>3.46-7.23</td>
</tr>
<tr>
<td>Multi chronic dx = n; High Dr. visits = y</td>
<td>Multi chronic dx = n; High Dr. visits = n</td>
<td>2.25</td>
<td>1.54-3.29</td>
</tr>
</tbody>
</table>

Table 4.4 High-cost users within moderate chronic health profile group, odds ratios (Total n = 112,776; High-cost users n = 11,277)

<table>
<thead>
<tr>
<th>Focus setting</th>
<th>Comparison setting</th>
<th>OR</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Home care (=yes)</td>
<td>Home care (=no)</td>
<td>2.04</td>
<td>1.94-2.14</td>
</tr>
<tr>
<td>Mental health (=yes)</td>
<td>Mental health (=no)</td>
<td>1.35</td>
<td>1.28-1.43</td>
</tr>
<tr>
<td>Hx of hosp = y ; low income = y</td>
<td>Hx of hosp = n; low income = y</td>
<td>4.76</td>
<td>4.32-5.23</td>
</tr>
<tr>
<td>Hx of hosp = y ; low income = n</td>
<td>Hx of hosp = n; low income = n</td>
<td>4.06</td>
<td>3.88-4.25</td>
</tr>
</tbody>
</table>
**Table 4.5** High-cost use within major newborn health profile group, odds ratio (Total n = 2,678; High cost users n = 267)

<table>
<thead>
<tr>
<th>Focus setting</th>
<th>Comparison setting</th>
<th>OR</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>History of hospitalization (=yes)</td>
<td>History of hospitalization (=no)</td>
<td>1.52</td>
<td>1.13-2.0</td>
</tr>
<tr>
<td>Low income = y; High Dr. visits = y</td>
<td>Low income = n; High Dr. visits = y</td>
<td>0.81</td>
<td>0.53-1.25</td>
</tr>
<tr>
<td>Low income = y; High Dr. visits = n</td>
<td>Low income = n; High Dr. visits = n</td>
<td>1.44</td>
<td>1.00-2.15</td>
</tr>
</tbody>
</table>

**Table 4.6** High-cost users within other mental health profile group (Total n = 62,381; High-cost users n = 6,241)

<table>
<thead>
<tr>
<th>Focus setting</th>
<th>Comparison setting</th>
<th>OR</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Home care = y; Hx of hosp = y</td>
<td>Home care = y; Hx of hosp = n</td>
<td>5.72</td>
<td>4.81-5.53</td>
</tr>
<tr>
<td>Home care = n; Hx of hosp = y</td>
<td>Home care = n; Hx of hosp = n</td>
<td>11.66</td>
<td>10.52-12.93</td>
</tr>
<tr>
<td>Low income = n; High Dr. visits = y</td>
<td>Low income = n; High Dr. visits = n</td>
<td>3.35</td>
<td>3.02-3.72</td>
</tr>
<tr>
<td>Low income = y; High Dr. visits = y</td>
<td>Low income = y; High Dr. visits = n</td>
<td>5.16</td>
<td>4.81-5.53</td>
</tr>
</tbody>
</table>

**Table 4.7** High-cost users within major acute disease health profile group (Total n=34,733; High cost users n=3,473)

<table>
<thead>
<tr>
<th>Focus setting</th>
<th>Comparison setting</th>
<th>OR</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length of hospital stay &gt;= 3 days</td>
<td>Length of hospital stay &lt; 3 days</td>
<td>35.5</td>
<td>26.8-47.1</td>
</tr>
<tr>
<td>Multi chronic dx = y; High Dr. visits = y</td>
<td>Multi chronic dx = n; High Dr. visits = y</td>
<td>2.1</td>
<td>1.9-2.3</td>
</tr>
<tr>
<td>Multi chronic dx = y; High Dr. visits = n</td>
<td>Multi chronic dx = n; High Dr. visits = n</td>
<td>3.8</td>
<td>2.3-6.2</td>
</tr>
</tbody>
</table>
Table 4.8 High-cost users within major cancer health profile group (Total n=8,711; High-cost users n=872)

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Odds ratio</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length of hospital stay &gt;= 3 days (=yes)</td>
<td>23.11</td>
<td>13.82-38.65</td>
</tr>
<tr>
<td>High emergency department visits (=yes)</td>
<td>1.59</td>
<td>1.36-1.85</td>
</tr>
<tr>
<td>History of hospitalization (=yes)</td>
<td>1.61</td>
<td>1.26-2.05</td>
</tr>
<tr>
<td>Home care (=yes)</td>
<td>2.36</td>
<td>2.01-2.78</td>
</tr>
<tr>
<td>Multiple (&gt;=3) chronic conditions (=yes)</td>
<td>1.68</td>
<td>1.44-1.97</td>
</tr>
</tbody>
</table>

Abbreviations:
- LOS = Length of hospital stay >= 3 days
- Multi chronic dx = Multiple (>=3) chronic conditions
- High Dr. visits = High physician visits
- Hx of hosp = History of hospitalization
4.5 Discussion

Health care systems in Canada and around the world are challenged with increased costs. The current study aimed to understand high-cost health care users from the perspective of a provincial health insurer. Using variables readily available in provincial administrative health databases, in combination with CIHI’s Population Grouping methodology, study results add to the evidence available to decision-makers as they develop policies to reduce costs, and, ultimately, improve the health of this population.

In the majority of health profile groups modelled, a measure of socio-economic status – neighbourhood income quintile – was statistically significantly associated with high-cost use.

Previous studies have also identified measures of socio-economic status (SES) associated with high-cost health care use, such as, unstable housing, and food insecurity, (Anderson et al., 2018; Rosella et al., 2014; Thavorn et al., 2017; Wrathall & Belnap, 2017), but, to the best of our knowledge this is the first population-based study to define high-cost users within their health profile group and consider SES in regression models. I hypothesize this method provides better specificity to understanding high-cost users of health care services.

Equity in health is of utmost importance. I feel that by providing quantitative evidence demonstrating the association between low socio-economic status and high-cost health care use within specific health profile groups – policy makers can create interventions aimed to both reduce costs and improve health.

A recent study out of the US similarly found that population segmentation methods were useful for defining actionable high-cost user cohorts (Joynt et al., 2017). The six cohorts the authors
defined were: under 65 years of age and disabled/end-stage renal disease; frail elderly; major complex chronic, minor complex chronic; simple chronic; and relatively healthy. Similar to the current study, individuals in the top 10% of spending were considered high-cost. The authors conclude that using simple criteria that segments, in this case, Medicare beneficiaries, into meaningful subgroups is a useful method to target interventions – a conclusion similar to the current study (Joynt et al., 2017). Besides only using data specific to one population (Medicare recipients), and, a non-validated approach to segmentation, this study is further limited by the fact that “high-cost use” was defined overall and not within each population segment. In addition, the authors did not conduct analyses beyond descriptive statistics, limiting the ability of their work to point to risk factors associated with high-cost use within the population segments.

Another recent study from the US also described the utility of population segmentation in describing high-cost users. This study employed density based cluster analysis to determine the population segments; however, their analysis did not include any indicators of socio-economic status (Powers et al., 2018). A recent systematic review of high-cost health care use, synthesizing 55 studies around the world, indicated that high-cost use was associated with multiple chronic conditions, older age, mental illness, end-of-life care, higher income (United States) and lower income (all other countries). Most relevant to the current work, however, the authors note that given the heterogeneity of high-cost user populations segmentation methods to define specific groups prior to analyses is recommended (Wammes et al., 2018).

Although our study population is limited to one provincial jurisdiction in Canada, the methods used can be generalized and used by other jurisdictions where a central health insurer is interested in defining their high-cost health care population. High-cost users are not a
homogenous group. Segmenting the population into health profile groups shows promise in describing the populations that consume the most health care resources.

Policy and decision-makers require actionable information. Providing descriptive epidemiology on high-cost health care users is not enough – to influence policy and decision-making actionable evidence must be provided. The methods and analyses conducted to arrive at a reasonable conclusion about what is driving high-cost health care use are often complex. It is incumbent upon implementation scientists to know how to communicate complex findings in a simple, easy to understand manner, in addition to focusing on what is amenable to change, such as socio-economic status, either at the individual or population-level.

4.5.1 Limitations

This study has several limitations, many inherent to epidemiological studies reliant on administrative health databases. Additional variables, such as food insecurity (Tarasuk et al., 2015) have been found to be associated with high-cost use and may have been useful in my modelling; however this data was not available for analysis. Community-based services, either publicly funded or private fee-for-service, (individual/group counselling, treatment centres, private nursing services, others) may be associated with high-cost use but not available for analysis. This is one of the largest drawbacks of studies making use of administrative health data – important confounders and predictors are often not available for analysis. Recent model simulation work has identified ways to impute some of these missing variables, though little is known if these methods work in practice (Silenou et al., 2019).

Readers will note, that the epidemiological context of this study is somewhat ‘muddy’ – the outcome of ‘high-cost use’ within specific health profile groups can be related to the definition
of the group itself (a combination of resource intensity and severity). This less than ideal context is acknowledged, however, I feel the population segmentation method still provides utility.

While I did account for the costs of prescription drugs, I did not examine the types of prescription drugs taken. There may well be an association between drug class and high-cost health care use.

Our study focused on cost; however, not all health care system costs were able to be measured, such as cancer treatment/diagnosis, laboratory testing, home care, public health, travel costs (air transfers and ground ambulance) and all health care administrative costs. High-cost users comprised a small proportion (10%) of the study population, yet accounted for 41% of measured health care costs. Given our inability to account for health care costs in previous studies, such as cancer diagnosis/treatment and laboratory costs, our findings are in general agreement with the literature.

In defining comorbidities, I relied on available health condition ‘tags’ in the Pop Grouper software; however, this method has not been validated against a gold standard method of chronic disease indicators (such as the Canadian Chronic Disease Surveillance System) (Feely et al., 2017). This study relied heavily on population segmentation, which may not always accurately classify individuals. For example, a person may be a high-cost user due to an opioid addiction, but, be classified into heart disease if the condition comprised their costliest health care utilization. Or, individuals who did have a specific health condition under study, but, did not ever seek medical treatment for it, would be classified as ‘health system non-users’. The segmentation methods may inadvertently not delineate specific groups associated with high-cost use, such as medical error (Powers et al., 2018).
I were unable to account for the specific days of insurance coverage for individuals who died/moved out of province; however, reassuringly 93% of study cohort members had complete follow-up time. As our study population comprised individuals followed over 24 months I did not consider high-cost health care use that persisted over time – differences in results would likely occur between episodic and persistent high-cost use.

Emergency department visits in Saskatchewan were limited by under-reporting as, during the study period, only <50% of the provincial emergency departments were NACRS-reporting facilities. Similarly, physician claims data would be limited in missing alternative-payment physician claims that do not shadow bill. Many psychiatrists in Saskatchewan are on alternative-payment schedules; this may disproportionately underestimate the number of individuals in mental health groups.

In terms of statistical analysis, it should be noted that all of the models, while demonstrating reasonable calibration and validation, had low pseudo $R^2$ values. This indicates the variables available for analysis are not explaining the majority of the variation between high-cost use and not high-cost use. The use of additional information not commonly available in administrative health databases, such as, social support, trust in the health care system, individual-level indicators of socio-economic status and others would likely greatly improve model variation explanation.

Costing data was missing for 54,948 individuals in the study population. It is possible that inclusion of those individuals would change model interpretation; however, given the small numbers of missing values (4%) I feel the impact of this limitation is likely to be minimal.
Additionally, high-cost health care use, while concerning from an insurer’s perspective, is not necessarily an adverse outcome. For example, high-cost health care use in obstetric patients may be entirely appropriate; driven by increased visits to specialists during a high-risk pregnancy and associated with better health outcomes. Quality of life outcomes were not assessed in this study.

Lastly, ‘high-cost’ use definitions vary and can be defined as individuals who exceed the 90th percentile, 95th percentile or 99th percentile of total population costs. I chose to focus on individuals in the top 10% of total costs per fiscal year as the majority of high-cost user studies in the literature (see Introduction) define ‘high-cost use’ as those in the top 10% of costs. The arbitrary nature of any cut-off value used, however, is acknowledged.

4.5.2 Future work

Our finding that ‘palliative’ is one of the costliest health profile groups is not unexpected (Aldridge and Kelly, 2015; Robst, 2015) and warrants further research to understand factors associated with this increased cost.

In only two health profile groups, ‘major acute’ and ‘major cancer’, high-cost use was not associated with lower socio-economic status. The factors associated with high-cost use in these two groups (major acute and major cancer) are health care utilization variables (such as lengthy hospital stays) which would be expected. It would be interesting to quantify, for major acute and major cancer patients, how many days in hospital could potentially be avoided by good continuity of health care providers in the community (less expensive use of resources compared to acute in-patient beds).
Future work could also consider high-cost use over time, potentially focusing on groups of trajectories, qualitative inquiries with high-cost users, linking health care costing data with other human services costing data to obtain a complete picture of ‘high-cost use’, and spatial analyses of high-cost use. The two costliest health profile groups in our study were palliative care and long-term care residents with the majority considered to be ‘high-cost users’. A detailed study into end-of-life costs and the reasons behind high-cost use for palliative patients (including specific interventions that not only save costs, but improve quality of life for patients); especially given previous contradictory findings regarding end-of-life costs (Aldridge & Kelley, 2015).

4.6 Conclusion

Given the heterogeneity, complexity and the natural ‘regression to the mean’ in health care spending, policy and decision-makers have difficulty devising interventions aimed at reducing high-cost health care use at a population level. By segmenting a provincial population into clinically meaningful sub-populations and demonstrating a link between socio-economic status and high-cost health care use in the majority of health profile groups, but not all; I feel this study adds to the body of evidence aimed at understanding the complexity of high-cost health care use.

Furthermore, this study provides quantitative evidence to support an agenda to improve equity in health. Within the majority of health profile groups, and taking into account a myriad of potentially confounding variables, individuals with lower socio-economic status were more likely to be high-cost health care users. Perhaps if policy makers addressed the underlying inequity, both improved population health and cost savings could be achieved.

By grouping the study population into health profiles, and understanding factors associated with high cost use within each specific health profile, the evidence generated becomes more specific,
and, potentially actionable. It is impossible to intervene at a system level on an individual’s age, but, system level interventions can be actioned and directed at improving socio-economic status, managing multiple comorbid conditions and improving end-of-life care.

There are relatively few examples of interventions that have successfully both reduced costs and improved health outcomes with respect to high-cost users. InterMountain Health Care in Utah and the Camden Primary Care consortium in New Jersey are good examples where, by focussing on the few consuming the most resources, cost savings and better health outcomes have been achieved. In both instances, high-cost individuals were a homogeneous group; in New Jersey the high-cost, high-utilization population was concentrated in a small geographical area and were of lower socio-economic status (Brenner, 2013). In Utah, the Connected Care Clinic was designed to serve the ‘complex few’ – those with multiple chronic conditions and multiple psycho-social issues (such as unstable housing and food insecurity). By ‘wrapping services around’ these complex few InterMountain was able to achieve positive results (Reiss-Brennan et al., 2010; Wrathall & Belnap, 2017).

Understanding the demographics, clusters, health care utilization patterns and predictors associated with high-cost health care use will be important for identifying opportunities for upstream prevention. This could be achieved through providing more targeted, appropriate care and supports for specific sub-populations, such as, mental health, newborns or multi-morbid individuals, or, by acting on the determinants of health to prevent certain types of high-cost use in the first place. For example, I found that for babies born in low income neighbourhoods, increased physician visits decreased the risk of high-cost use. Related policies could include identifying at risk pregnant mothers and connecting them to primary care providers prior to the birth of their baby. Similarly, home care services could be a priority for mental health patients.
coming out of hospital. Given that inequity was related to high-cost use within most health profile groups, disadvantaged/vulnerable persons with major chronic disease, for example, could be provided with one-on-one social worker support – decreasing health care costs and improving quality of life. I feel that by developing well calibrated and discriminatory models aimed at understanding factors associated with high-cost use I am providing a piece of the puzzle for policy makers keen to implement interventions. If successful, these interventions could both lower costs, but more importantly, improve the health of the population.
4.7 References


https://doi.org/10.1258/13558190360468164


https://doi.org/10.1186/s12913-014-0532-2


https://doi.org/10.1371/journal.pone.0211118


4.8 Supplemental material

4.8.1 CIHI Population Grouping Methodology

Individuals rarely have a single health condition. At the request of the Canadian Council of Deputy Ministers of Health, CIHI was tasked with creating a method that could reliably group provincial populations into clinically meaningful sub-groups. CIHI undertook this work over the course of eight years creating expert scientific working groups for each condition, including leading Canadian clinical experts, in addition to an overall scientific steering committee made up of researchers, academics, policy and decision-makers. Using a combination of administrative health databases and provincial health registry systems, the Population Grouping Methodology (‘Pop Grouper’) ‘tags’ each resident with any of the 239 health conditions. These binary tags (0/1) form the ‘building block’ of the grouping methodology and are not mutually exclusive; an individual can have any number of applicable health conditions. The presence of health conditions are determined by linking data from hospitalizations, physician visits, hospital day surgeries, emergency department visits and long-term care. Pop Grouper uses 24 months of data to determine health conditions (n = 239), branches (n = 164) and health profile groups (n = 16).

Groups of health conditions that are clinically related are addressed through the use of ‘clinical overrides’. These overrides were developed by each of the clinical scientific working groups. Overrides were designed to reduce redundancies with the following rules: 1) More severe manifestations of disease override less severe, 2) Overrides ‘fix’ differences in provincial coding practices (for example, ‘influenza’ will override ‘unspecified infection’, and, 3) Diagnosis will override symptoms (for example, acute gastrointestinal hemorrhage will override abdominal pain when concurrent). The following depicts how clinical overrides work using cardiovascular disease grouping as an example.
Specific health condition tagging rules were applied to physician billing data. Physician billing data in Canada has known limitations, largely due to the ability to code for only one three digit diagnosis, and, the fact that this data source is not subject to rigorous, national data quality initiatives. For some conditions identified through physician data only, the condition needs to occur more than once over the 24 month period in order to ‘count’ as a health condition. For example, for the ICD-9 code 300, ‘anxiety’, if coded in the physician billing data would only ‘tag’ an individual with the anxiety health condition if it occurred >= 2 times within the 24 month period in the physician data, and, never in any other data sources. The purpose of these physician data only rules was to minimize false positive health conditions. All physician data
rules fall into one of the following categories, based on specific diagnosis: 1) one diagnosis needed to ‘count’, 2) \( \geq 2 \) physician visits on two or more separate dates, or, 3) no diagnosis needed (for example, ‘healthy newborn’ assigned based on age and/or hospitalization, no physician visit is required to ‘count’).

Next, each of the 239 health conditions are compiled into 164 branches ranked in order from most severe to least severe. Severity ranking was conducted by each technical working group, including clinicians. The following criteria were used when assessing level of severity, 1) clinical complexity and 2) resource intensity. These criteria were assessed through clinical judgement and average costing data. The branches compile related health conditions into the same ‘branch’; branches often split based on the presence or absence of significant comorbid conditions. ‘Significant’ comorbid conditions were determined through descriptive data analyses with input from the Pop Grouper clinical expert working groups; assessment consisted of two categories: 1) If sufficient numbers of individuals with or without comorbid conditions existed, and, 2) If there was a substantial difference in cost between individuals with or without comorbid conditions. Similar to health conditions, branches are not mutually exclusive.

Each of these branches is, in turn, linked to one of 16 mutually exclusive health profile groups. Where an individual belongs to more than one branch the one with the largest severity/resource intensity is used for health profile group determination. Health profile group severity, as with health conditions, was determined based on a combination of clinical judgement and resource intensity. The graphic below from CIHI lists each of the mutually exclusive health profile groups in ranked order of severity (palliative care group as the most severe; health care non-users as least severe).
As the following studies focus on health care cost as an outcome it is important to discuss costing data in the Grouper, and, the costing methods used in this thesis. Pop Grouper does contain costing data, however, I chose to not use Pop Grouper costing data and instead use the person-level costing method and population threshold values described previously.

The reasons for this are three-fold: 1) for the sake of consistency of results, 2) person-level costing methods are validated but Pop Grouper costing data is not yet validated, and, 3) in descriptive analyses of Pop Grouper costing data I was concerned about the ‘circular’ use of the costing function in Pop Grouper; health profile groups are assigned, at least partially, on resource intensity. The limitations to the Pop Grouper costing method are, in general, due to the limitations of CIHI’s data holding compared to provincial data holdings. The Pop Grouper costing data limitations are: 1) extensive cost imputation methods are used for data gaps, and, 2) no data was available for long-term care or home-care costs in Pop Grouper. Both of these data limitations were greatly reduced by using the Saskatchewan databases I had access to within...
province. As a result, all three studies detailed below used the person-level costing data to determine individuals who were ‘high-cost’ in all three studies.

### 4.8.2 Effect estimates

Odds ratios values are reported in ‘Results’; all effect estimates by health profile group follow.

**Table 4.8.2.1: High-cost use within major chronic health profile group, effect estimates (Total n = 37,803; High-cost users n = 3,781)**

<table>
<thead>
<tr>
<th>Effect</th>
<th>β</th>
<th>SE β</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Home care (=yes)</td>
<td>0.67</td>
<td>0.04</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Length of hospital stay ≥ 3 days (=yes)</td>
<td>2.73</td>
<td>0.09</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Low income (=yes)</td>
<td>-0.50</td>
<td>0.23</td>
<td>0.134</td>
</tr>
<tr>
<td>Length of hospital stay ≥ 3 days X Low income (yes, yes)</td>
<td>0.66</td>
<td>0.23</td>
<td>0.004</td>
</tr>
<tr>
<td>High physician visits (=yes)</td>
<td>0.82</td>
<td>0.19</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Multiple (≥3) chronic conditions (=yes)</td>
<td>-0.20</td>
<td>0.26</td>
<td>0.144</td>
</tr>
<tr>
<td>High physician visits X Multiple (≥3) chronic conditions (yes, yes)</td>
<td>0.80</td>
<td>0.27</td>
<td>0.003</td>
</tr>
<tr>
<td>Intercept</td>
<td>-5.90</td>
<td>0.19</td>
<td></td>
</tr>
</tbody>
</table>

Pseudo $R^2 = 0.27$; Hosmer and Lemeshow Goodness-of-Fit: $\chi^2=14.4$; df=8; $p=0.1$
Table 4.8.2.2: High-cost users within moderate chronic health profile group, effect estimates (Total n = 112,776; High-cost users n = 11,277)

<table>
<thead>
<tr>
<th>Effect</th>
<th>β</th>
<th>SE β</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Home care (=yes)</td>
<td>0.71</td>
<td>0.03</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Mental health condition (yes)</td>
<td>0.30</td>
<td>0.03</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>History of hospitalization (yes)</td>
<td>1.40</td>
<td>0.02</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Low income (yes)</td>
<td>-0.10</td>
<td>0.04</td>
<td>0.510</td>
</tr>
<tr>
<td>History of hospitalization X Low income (yes, yes)</td>
<td>0.16</td>
<td>0.05</td>
<td>0.003</td>
</tr>
<tr>
<td>Intercept</td>
<td>-2.78</td>
<td>0.02</td>
<td></td>
</tr>
</tbody>
</table>

Pseudo $R^2$ = 0.10; Hosmer and Lemeshow Goodness-of-Fit: $\chi^2$=8.3; df=4; $p$=0.1
Table 4.8.2.3: High-cost use within major newborn health profile group, effect estimates (Total n = 2,678; High cost users n = 267)

<table>
<thead>
<tr>
<th>Effect</th>
<th>β</th>
<th>SE β</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>History of hospitalization (=yes)</td>
<td>0.41</td>
<td>0.15</td>
<td>0.005</td>
</tr>
<tr>
<td>Low income (=yes)</td>
<td>0.37</td>
<td>0.20</td>
<td>0.602</td>
</tr>
<tr>
<td>High physician visits (=yes)</td>
<td>1.23</td>
<td>0.17</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Low income X High physician visits (=yes, yes)</td>
<td>-0.58</td>
<td>0.30</td>
<td>0.054</td>
</tr>
<tr>
<td>Intercept</td>
<td>-2.91</td>
<td>0.12</td>
<td></td>
</tr>
</tbody>
</table>

Pseudo $R^2$=0.10; Hosmer and Lemeshow Goodness-of-Fit: $\chi^2=3.23$; df=5; $p=0.67$
Table 4.8.2.4: High-cost users within other mental health profile group, effect estimates (Total n = 62,381; High-cost users n = 6,241)

<table>
<thead>
<tr>
<th>Effect</th>
<th>$\beta$</th>
<th>SE $\beta$</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Home care (=yes)</td>
<td>0.90</td>
<td>0.07</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Low income (=yes)</td>
<td>0.52</td>
<td>0.05</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>History of hospitalization (=yes)</td>
<td>2.46</td>
<td>0.05</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>High physician visits (=yes)</td>
<td>1.64</td>
<td>0.04</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Home care X History of hospitalization (yes, yes)</td>
<td>-0.71</td>
<td>0.18</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Low income X High physician visits (yes, yes)</td>
<td>-0.43</td>
<td>0.06</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Intercept</td>
<td>-3.25</td>
<td>0.03</td>
<td></td>
</tr>
</tbody>
</table>

Pseudo $R^2$ = 0.15; Hosmer and Lemeshow Goodness-of-Fit: $\chi^2$=9.4; df=4; $p$=0.1
Table 4.8.2.5: High-cost users within major acute disease health profile group, effect estimates (Total n=34,733; High cost users n=3,473)

<table>
<thead>
<tr>
<th>Effect</th>
<th>β</th>
<th>SE β</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length of hospital stay ≥ 3 days (=yes)</td>
<td>3.57</td>
<td>0.14</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>High physician visits (=yes)</td>
<td>1.84</td>
<td>0.17</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Multiple (≥3) chronic conditions (=yes)</td>
<td>1.32</td>
<td>0.26</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Multiple (≥3) chronic conditions X High physician visits (yes, yes)</td>
<td>-0.59</td>
<td>0.26</td>
<td>0.025</td>
</tr>
<tr>
<td>Intercept</td>
<td>-7.39</td>
<td>0.21</td>
<td></td>
</tr>
</tbody>
</table>

Pseudo R²=0.26; Hosmer and Lemeshow Goodness-of-Fit: χ²=0.6; df=4; p=0.97

Table 4.8.2.6 High-cost users within major cancer health profile group (Total n=8,711; High-cost users n=872)

<table>
<thead>
<tr>
<th>Predictor</th>
<th>β</th>
<th>SE β</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length of hospital stay ≥ 3 days (=yes)</td>
<td>3.14</td>
<td>0.26</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>High emergency department visits (=yes)</td>
<td>0.46</td>
<td>0.08</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>History of hospitalization (=yes)</td>
<td>0.48</td>
<td>0.13</td>
<td>0.0001</td>
</tr>
<tr>
<td>Home care (=yes)</td>
<td>0.86</td>
<td>0.08</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Multiple (≥3) chronic conditions (=yes)</td>
<td>0.52</td>
<td>0.08</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Intercept</td>
<td>-6.16</td>
<td>0.26</td>
<td></td>
</tr>
</tbody>
</table>

Pseudo R²=0.25; Hosmer and Lemeshow Goodness-of-Fit: χ²=7.01; df=8; p=0.54
Cohort creation for Chapter 4, compared to Chapters 2 and 3, proceeded differently as I wanted to focus on the entire provincial population, not restricting study data to mental health and addictions patients only. In conducting the analyses for Chapters 2 and 3 I became interested in how I could potentially segment the provincial population into clinically meaningful and distinct sub-groups to understand high-cost health care use in general. The Canadian Institute for Health Information released a newly developed Population Grouping Methodology (beta release: April 2017; version 1.0: April 21, 2018) which is similar to diagnosis related grouping methods available elsewhere but with no additional cost to collaborating provinces and territories. This method proved not only useful for my aim of segmenting the provincial population, but, also for study cohort creation.

The cohort under study in Chapter 4 constitutes all Saskatchewan residents with a valid health services card with at least 1 day of coverage from April 1 2015 to March 31 2017, including health system non-users (‘study cohort’). The Population Grouping (‘Pop Grouper’) Methodology developed by the Canadian Institute for Health Information and modified for use with Saskatchewan datasets (see 1.3.5 for details) was ran to create this study cohort. In Saskatchewan, the Pop Grouper exists as hundreds of lines of SAS code, created and validated by CIHI and modified for Saskatchewan specific database and variable names by in-province analysts. The code requires access to raw administrative health databases and ~24 hours to run. In Saskatchewan, Pop Grouper SAS code has been independently run and results validated by the Canadian Institute for Health Information, the Saskatchewan Ministry of Health and the Saskatchewan Health Quality Council.
Health care costing data extraction proceeded as described in section 2.7.2; calculations of total study cohort and population threshold values differed (see 4.7.2).

4.8.3.1 Health care utilization and data manipulation

All health care utilization data (physician visits, emergency department visits, long-term care residency, and hospitalizations) were pulled from raw data files for study cohort members as separate files per databases in ‘long’ format. I manipulated this data using ‘do loops’, ‘arrays’ and ‘transpose’ statements to create the variables I wanted to explore in the analyses.

4.8.3.2 Hospital data

In the hospital data, I first excluded all hospital transfers as they are considered to be ‘one’ hospitalization and multiple transfers can occur during one episode of care. Transfers were identified through a validated ‘episode of care’ variable created for research purposes by the Saskatchewan Health Quality Council. I calculated length of stay by subtracting the admission date from the discharge date. All hospitalizations and lengths of stay were summed by fiscal year per person and transposed to ‘wide’ format; this created a data set with one record per study cohort member.

4.8.3.3 Physician billing data

In the physician data, I first excluded all ‘slush’ codes. This a group of specific physician codes that relate to out-of-province visits, telephone consults, insurance form completion, etc. and are typically excluded from research studies. Physician visits that occurred on the same day with the same physician were considered duplicate and excluded. Physician visits were summed per fiscal
year per person, using arrays, do loops and transpose statements, resulting in a data set with one record per study cohort member.

4.8.3.4 Emergency department

Emergency department data was fully complete for Chapters 2 and 3, as this study cohort focused on Saskatoon only. For Chapter 4, emergency department visits occurring in non-NACRS reporting facilities (see 1.3.1.2 for a description of NACRS) were not available for analysis. Emergency department visit data was manipulated similarly to physician and hospital data – summing the number of visits per fiscal year per study subject. Multiple visits by the same person on the same day, where the time stamp of arrival differed, were considered more than one visit.

4.8.3.5 Health care costing data

Using the health care costing data files for the provincial population of Saskatchewan I conducted two analyses. First, I calculated the population threshold value for each fiscal year (2015/16 and 2016/17) – this was the dollar amount associated with the 90\textsuperscript{th} percentile in total health care costs. Second, I extracted total health care costs, by service used, for fiscal years 2015/16 and 2016/17, for each member of the study cohorts. Any study cohort member who exceeded the 90\textsuperscript{th} percentile of total health care costs at the population level (population threshold value) was considered to be ‘high-cost’.

Using the date of death, I calculated the number of days per fiscal year each study subject was alive and able to access health services. Any individual experiencing a break in health service coverage and may have moved in or out of the province in the 24 months under study were
included; however, due to a technical limitation I was unable to determine the total number of days each individual was covered. For this reason, total costs, both study cohort members and population threshold values, were calculated per fiscal year (2015/16 and 2016/17).

4.8.3.6 Long-term care and home care data

Long-term care and home-care data were linked at the individual-level for each study cohort member. Total home-care visits by fiscal year were derived. Long-term care flags (Yes/No) per fiscal year were created for each study cohort member. As the analyses progressed it became clear that LTC residents were nearly entirely persistently high-cost given the cost of a LTC stay; starting in Chapter 3 all LTC residents were excluded from subsequent analyses.

Comorbidities were defined using the ‘health condition’ binary flags in the Population Grouping Methodology. Any study cohort member with >= 2 chronic health conditions were considered to be ‘multi-morbid’.
4.8.4 Statistical modeling

In Chapter 4, after employing the CIHI Population Grouping methodology, I create separate logistic regression models for each of the mutually exclusive health profile groups. I defined high-cost use within the health profile groups – sensitivity analyses demonstrated that large differences in co-efficients and interpretation occurred when defining high-cost in the study population overall versus within the sub-groups. As the aim of the study was to understand risk factors associated with high-cost use within clinically meaningful sub-populations I include the models where high-cost use was defined within the groups. As opposed to Chapter 2, I found that using this definition of high-cost use improved model fit to the point that I could employ logit links, as opposed to probit. Model building efforts proceeded similarly to Chapter 2 (assessment of confounding, biologically meaningful interactions and incremental loss/gains in AIC).
4.8.5 Co-author affiliations and contributions

Maureen Anderson MSc, PhD (candidate)\textsuperscript{1,2*}, Crawford W Revie MSc, PhD\textsuperscript{1,3}, Cordell Neudorf MD, MHSc\textsuperscript{1,5}, Yvonne Rosehart BA\textsuperscript{6}, Wenbin Li MSc, PhD\textsuperscript{2}, Meriç Osman MA\textsuperscript{2}, David L. Buckeridge MD, MSc, PhD\textsuperscript{7}, Laura C Rosella MSc, PhD\textsuperscript{8-11}, Henrik Stryhn PhD\textsuperscript{1}, Walter Wodchis MA, MAE, PhD\textsuperscript{10,11}

\textsuperscript{1} Department of Health Management, University of Prince Edward Island, Charlottetown, Prince Edward Island, Canada
\textsuperscript{2} Saskatchewan Health Quality Council, Saskatoon, Saskatchewan, Canada
\textsuperscript{3} Department of Computing and Information Sciences, University of Strathclyde, Glasgow, Scotland
\textsuperscript{4} Department of Community Health and Epidemiology, College of Medicine, University of Saskatchewan, Saskatoon, Saskatchewan, Canada
\textsuperscript{5} Population and Public Health, Saskatchewan Health Authority, Saskatoon, Saskatchewan, Canada
\textsuperscript{6} Canadian Institute for Health Information, Ottawa, Ontario, Canada
\textsuperscript{7} Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, Quebec, Canada
\textsuperscript{8} Public Health Ontario, Toronto, Ontario, Canada
\textsuperscript{9} Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada
\textsuperscript{10} Institute for Clinical Evaluative Sciences, Toronto, Ontario, Canada
\textsuperscript{11} Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, Canada

\textbf{MA} – Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Software, Validation, Visualization, Writing: original draft preparation, editing, Final approval and Agreement to be accountable

\textbf{CR} – Conceptualization, Methodology, Visualization, Supervision, Writing: review and editing, Final approval and Agreement to be accountable

\textbf{CN} – Conceptualization, Funding Acquisition, Resources, Supervision, Writing: review and editing, Final approval and Agreement to be accountable

\textbf{YR} – Methodology, Validation, Writing: original draft preparation, review and editing, Final approval and Agreement to be accountable

\textbf{WL} – Data Curation, Formal Analysis, Investigation, Software, Writing: review, Final approval and Agreement to be accountable

\textbf{MO} – Data Curation, Formal Analysis, Investigation, Methodology, Software, Validation, Writing: review and editing, Final approval and Agreement to be accountable

\textbf{DB} – Conceptualization, Validation, Supervision, Final approval and Agreement to be accountable
LC – Conceptualization, Methodology, Writing: review, Final approval and Agreement to be accountable

HS – Investigation, Methodology (statistics), Writing (review), Final approval and agreement to be accountable

WW – Conceptualization, Funding Acquisition, Methodology, Supervision, Writing: review, Final approval and Agreement to be accountable
5. CONCLUSION

5.1 Summary of main findings

In an age of increased health care costs and budgetary pressures at all levels of government, continuing to deliver universal health care is a challenge. It is compelling that, without exception, in every context studied (small centres, large urban centres, entire provinces and entire countries) a small proportion of the population consumes the majority of health care resources.

A recent systematic review of the literature of high-cost patients conducted by Wammes et al focussed on 55 studies around the world in an effort to compare high-cost users across countries and payment models. The authors conclude there are similarities in this population globally and most often consist of patients with many chronic conditions and mental health and addiction issues. The majority of the studies found were conducted in the US and Canada. Interestingly, in the US high-cost health care users were more likely to have higher incomes, but, the opposite was found in all other countries reviewed (Wammes et al., 2018).

In a 2016 New England Journal of Medicine article, Dr. David Blumenthal and colleagues state that finding better care models for high-cost, high-need patients is an urgent priority (Blumenthal, Chernof, Fulmer, Lumpkin, & Selberg, 2016). It thus seems reasonable to focus efforts on understanding this population in order to reduce costs, but more importantly for me, to improve the health of the population.

The studies in this dissertation have contributed novel findings to the, relatively recent, body of literature surrounding high-cost users of health care resources. In Chapter 2, I found that the average health care cost among study cohort members was ~ $2,300/year; while for high-cost
users it was ~$19,000. In this study I focused on persistent high-cost users among a cohort of mental health and addiction clients – those who were persistently in the top decile of spending. In logistic models, individuals with unstable housing and hospitalization(s) had increased risk of persistent high-cost use; both of these effects were more pronounced as comorbidities increased. Patients with schizophrenia, particularly those under 50 years old, had an increased likelihood of persistent high-cost use. The probability of persistent high-cost use decreased with good connection to a primary care provider; this effect was more pronounced as the number of mental health conditions increased. These findings provide an anchor for policy-makers interested in factors associated with persistent high-cost use (comorbid conditions, severe mental illness and unstable housing) and a factor that I found to be protective of high-cost use (good connection to a primary care provider).

In Chapter 3, within the same study population as Chapter 2, mental health and addiction clients, (n = 124,497), I found that the majority (~65%) of cohort members were ‘never high-cost’. Persistent high-cost users, those that were high-cost year after year, represented a small component of the study population (n = 2,352; 2%) and accounted for 15% total health care costs. This chapter focused on predicting those at risk of future high-cost use – ideally providing a potential point of intervention prior to becoming high-cost. My findings point to, within the study population, those variables that are most important for predicting one-year and persistent high-cost use. Classification and Regression Trees were used to visualize where ‘splits’ occurred in variables leading to higher proportions of high-cost users. I found both episodic and persistent high-cost use the number of physician visits in previous years can be used as a predictor – a count that is easily obtained by central health insurers and/or provincial health departments.
Applying innovative analytical techniques, including prediction, to both common (for example, administrative health data) and less common (for example, social media data) data sources may provide important evidence to inform health policy and intervention. The use of predictive algorithms in marketing is not new; the use of the same in health care has, arguably, reached less than its full potential (Yeung, 2018). Individuals who are high-cost and/or high utilizers of health care can flag points in the medical system where it is failing. If I can predict, with reasonable discrimination and calibration, those at risk of future high-cost use then cost savings can result, but, perhaps more importantly, opportunities exist to improve health outcomes and the experience of the patient.

In Chapter 4, within a general provincial cohort (n = 1,175,147), I found that high-cost users consumed 41% of total health care resources. Average annual health care spending for individuals who were not high-cost was ~ $650; while for high-cost users the figure was over $16,000. The costliest health profile groups were ‘long-term care’, ‘palliative’, ‘major acute’, ‘major chronic’, ‘major cancer’, ‘major newborn’ and ‘major mental health’. The majority of the high-cost user population were grouped into chronic disease (n = 24,950), acute disease (n = 17,511), obstetrics (n = 12,794), and mental health (n = 6,581). One measure of socio-economic status, low neighbourhood income, was not statistically significantly associated with high-cost use in the health profile groups of ‘major cancer’ or ‘major acute’. However, there was a statistically significant association between lower socio-economic status and high-cost health care use in the following health profile groups: ‘major chronic’, ‘moderate chronic’, ‘major newborn’ and ‘other mental health’. Unfortunately, due to a systematic error in ICD coding in Saskatchewan administrative health databases the health profile group, ‘major mental health’ was not available for analysis. Neighbourhood income level is not a perfect indicator of socio-
economic status, but happens to be data that is readily available in administrative health
databases. An earlier Canadian study found good agreement between the ecological indicator at
the neighbourhood level and individual household wealth (Mustard, Derksen, Berhelot, &
Wolfson, 1999).

These findings emphasize the need for grouping heterogeneous populations into homogeneous
groups before developing models aimed at understanding the outcome of interest. By grouping
individuals into relevant health profile groups and then defining high-cost use within each group
I feel I have created a clearer understanding of the factors associated with high-cost health care
use.

The overall findings of the studies within this thesis are, reassuringly, not incongruent with a
recent systematic review on high-cost health care use. This review found high-cost users were
more likely to have mental health and addictions, multiple chronic conditions, and, in countries
outside of the US, lower incomes. The results in the preceding Chapters found the same results,
however, I would argue, take the understanding of high-cost use a step further.
The findings presented here provide a greater level of specificity, from the perspective of a
provincial health care insurer, to the high-cost user phenomenon. I feel this can best be
summarised in the following four points:

1) By segmenting the population into clinically meaningful sub-groups, and importantly,
defining ‘high-cost use’ within each sub-group, my findings demonstrated that low
income was associated with the majority of high-cost groups, but not all. This finding
could be useful for health insurers looking to intervene on the social determinants of
health and providing actionable information leading to particular clinical sub-groups that would be worth the targeted effort. Further, this finding may make decision-makers more apt to believing the findings – reassuring them that socio-economic status is not associated with high-cost use in all clinical sub-populations, but action can be targeted where it is needed most.

2) I quantitatively demonstrated a protective factor to high-cost health care use – having a good connection to a primary care provider. This finding is positive, and, one that can be used by primary care practitioners in advocacy work; in addition to health care policy makers arguing for provincial investment in primary care.

3) Predicting individuals at risk of future high-cost use is not technically difficult. I demonstrated that, by using routinely available administrative health data today, individuals at risk of high-cost use in the future can be flagged for intervention with a high degree of specificity. Health management organizations in the US use predictive algorithms and have demonstrated success in intervention programs (see section 5.2); however, in Canada, with few exceptions, I do not make enough use of the routinely collected administrative health data, not only for descriptive statistics, but, for conducting analytic modelling.

4) As evident by the use of the ‘shared client index’, the ‘street nurse outreach data’ and other data sources I took advantage of for these studies, there are powerful databases with relevant information that exist in disparate organizational silos. The process from my looking for a way to quantify, at an individual level, measures of socio-economic status in administrative health data; to actually having access to the data took nearly 3 years. I was in a fortunate position of working full-time as an epidemiologist while
simultaneously completing this doctoral thesis; but for many students, arguably the folks in the system tasked with conducting this type of work, do not have the luxury of the vast amount of time it takes to jump through hurdles of privacy impact assessments, data sharing schedules, collaborative agreements between many parties and varying levels of government, extracting data from systems never before used for research, de-identification processes and many more. This phenomenon, unfortunately, is not unique to Saskatchewan. There are other Canadian jurisdictions do have better systematic processes in place for data access (such as the Institute for Clinical Evaluative Sciences in Toronto, or, the Manitoba Centre for Health Policy in Winnipeg), but, the time constraints and administrative hurdles, despite being lessened, still exist. It is a poor state of affairs that private industry uses ‘big data’ and predictive algorithms from multiple sources to target advertisements for shoes I may like (which I can personally attest to the high degree of accuracy of the algorithms they are running) more frequently and to a better degree than in health care. Canada needs a comprehensive electronic medical health record that contains all health data for all people, accessible to all health care providers, and, in the case of our individual health history, to patients themselves. There are models from other jurisdictions that have demonstrated that this possible (Denmark, Sweden), while still maintaining appropriate privacy controls and data access permission by health care provider role. In 2019 it is inexcusable that residents of Canada cannot log into a secure web-based application to access their own health information.
5.2 High-cost user research: interventions to reduce cost and improve health

There are relatively few examples of interventions that have successfully both reduced costs and improved health outcomes with respect to high-cost users. InterMountain Health Care in Utah and the Camden Primary Care consortium in New Jersey are good examples where, by focussing on the few consuming the most resources, cost savings and better health outcomes have been achieved. In both instances, the high-cost individuals were a homogeneous group; in New Jersey the high-cost, high-utilization population was concentrated in a small geographical area and were of lower socio-economic status (Brenner, 2013). In Utah, the Connected Care Clinic was designed to serve the ‘complex few’ – those with multiple chronic conditions and multiple psycho-social issues (such as unstable housing and food insecurity). By ‘wrapping services around’ these complex few InterMountain was able to achieve positive results (Reiss-Brennan et al., 2010; Wrathall & Belnap, 2017).

A systematic review of high-cost, high-need interventions by Bleich et al categorized two models of intervention: 1) case/care management and 2) disease management. In case/care management, the program worked to provide better continuity of care between health care providers and/or assisting the individual client with their unique circumstances requiring care. The first type of model generally demonstrated a reduction in health care utilization. Disease management models of intervention, however, focused on the chronic disease and worked to control it; which generally resulted in better health outcomes (Bleich et al., 2015).

Interestingly, the quintuple aims of health care – reduce costs, improve population health, improve patient experience, improve provider experience and improve health equity (Bodenheimer & Sinsky, 2014; Feely, 2017) – can be linked to interventions aimed at high-cost,
high-need populations. From Bleich and colleagues’ review article, I find that, in general, disease management models improve health outcomes, while case management models reduce costs. This is the conundrum of the quintuple aim – achieving one of the first two aims with any given intervention is relatively easy; what is hard to achieve is both reduced costs and improved health outcomes at the same time.

A qualitative investigation comparing the operational approaches of 18 successful programs geared to high-cost, high-need patients was conducted by Hong et al. The authors conclude, “effective programs customize their approach to their local contexts and caseloads; use a combination of qualitative and quantitative methods to identify patients; consider care coordination one of their key roles; focus on building trusting relationships with patients as well as their primary care providers; match team composition and interventions to patient needs; offer specialized training for team members; and use technology to bolster their efforts.” (Hong, Siegel, & Ferris, 2014).

The majority of high-cost user published studies have been conducted in North America. A recent systematic review of 55 high-cost health care user studies across the world indicated that relatively few of those studies were conducted outside of the United States and Canada (see 1.3.4 for a detailed summary of the systematic review) (Wammes et al., 2018).

5.3 High-cost user research: challenges and lessons learned

The challenges in researching high-cost health care users are many. The population of high-cost users at any given time will be different. I would encourage other researchers, policy-makers and
decision-makers, when working to tackle this issue to approach the analyses in a concerted fashion and would offer some specific insights based on the work presented within this thesis.

First, use solid, evidence-based approaches that adopt a ‘bottom-up’, person-level costing method, such as those developed by health economists in Canada (Woodchis et al., 2013), to attribute health care costs at the individual level. This method provides better specificity to health care costing data at the individual level.

Second, divide the population of interest, using more than one year of costing data, into ‘episodic’ high-cost use and ‘persistent’ high-cost use. I know from published estimates, and from my results in Chapter 2 and 3, that approximately 30% of high-cost users will go on to be high-cost the following year (Wodchis, Austin, & Henry, 2016). Episodic and persistent high-costs are different and the interventions that can mitigate their occurrence will also likely be different.

Thirdly, from findings in Chapter 4, and recently in the United States (Joynt et al., 2017), I suggest that high-cost user research should first segment the population under study into homogenous groups in order to better understand risk factors associated with high-cost use and delineate, as best one can, between ‘appropriate’ high-cost use (for example, catastrophic illness, trauma or long-term care resident) and potentially more modifiable risk factors (for example, socio-economic status). These examples, I feel, demonstrate the importance of parsing out high-cost, high-utilization populations into homogenous sub-groups. It is not enough to simply sort a provincial client registry into the top 5% or top 10% of spenders and intervene on an unanalyzed sub-set. First, I would argue, analyses must be conducted to group ‘high-cost users’ into homogeneous groups, before I can better understand the factors underlying such high-cost use.
As an illustrative example of the importance of creating homogenous groups, I found consistently in all three of my studies (as did Robst et al), that long-term care residents are high-cost users (Robst, 2015). Is this a homogenous population where interventions can reduce costs? Perhaps, but not likely; the majority of the health care costs incurred by those thus institutionalized are due to the costs of institutionalization itself. One long-term care bed in Saskatchewan costs up to $2,700 per year – at or above the threshold of high-cost health care use (Saskatchewan Ministry of Health, 2017).

The largest limitation and constraint to the studies presented in this thesis relate to the data source used. Administrative health data that is routinely collected within the health care system is not intended for research. As research is a by-product of this administrative data, it is impossible to categorize all of the individual-level risk factors that could be associated with high-cost use. For example, the quality and amount of social support, food security, adverse childhood events, the level of trust an individual has in their primary care provider, measures of obesity, physical activity, residential school survivorship, educational attainment and a great many more potential predictors of a high health care cost outcomes are impossible to analyze at the population level. The influential 1974 Lalonde report (Canadian) was an influential document around the world outlining the role of these non-health care system factors, and specifying them. Perhaps, as ‘big data’, digital technology and mobile data collection increase in prevalence researchers will be able to better understand this information at an individual level; incorporating, possibly, variables that ‘explain’ the variance in exploratory models better, such as level of trust in the health care system, than routinely collected administrative health data.
5.4 Future work

Future studies, I feel, would benefit from taking these analyses one step further. It would be useful to conduct a robust cost-benefit or return on investment analysis; using all government costs, not just health service costs. Future work should attempt to attribute costs to individuals in lower socio-economic status populations compared to individuals in higher socio-economic status circumstances. Successful program interventions aimed at addressing poverty could be costed and compared to the amount of government spending that is attributed to individuals of low socio-economic status. It would be interesting to see findings from this type of analysis at a population level. Anecdotal evidence from front-line health care workers (street nurses and social workers) in Saskatoon reinforced my study findings. For example, a social worker relayed a story where, for the small cost of a promised buffet lunch at the end of the month, one individual’s twice daily visits to the Emergency Department were avoided. These stories are compelling; however, I feel the findings need to be evident at a population-level with data-driven analyses, including government costs, in order to provide the best evidence possible for decision-making.

The health care system in Saskatchewan is in the midst of implementing a concept of health networks; these ‘teams’ of providers assigned to specific geographical boundaries are to work collaboratively together to address the health needs of the population. I do not claim that programs geared towards high-cost, high-need patients need to be created and take away from system resources aimed at primary care transformation. Instead, I would challenge health care decision-makers to ensure that the small proportion of the population consuming the majority of health care resources are not neglected in primary care transformation planning and left to
navigate an increasingly complex health care system on their own. In fact, I suggest that decision-makers include representative high-cost patients as part of primary care transformation project teams and that those interventions need to consider socio-economic status.

Using county-level data from the United States, it has been estimated that health care services influence only 20% of an individual’s health status (County Health Rankings, 2015); though recent work has challenged the simplicity of the mathematical formulas behind these calculations and that it is erroneous to assume that health can be summed to 100% (Krieger, 2017). But regardless of the specific proportion attributed, it is well accepted that health care services are but a small part of overall health and wellbeing. The majority of health status is determined by things outside of health: income, education, adverse childhood events, social support, physical activity, and others. As part of its constitution, the World Health Organization states ‘health’ is more than merely the absence of disease (World Health Organization, 2015). However, health care systems have been slow to adapt and change policies to incorporate ‘health as all policy’ system wide. A recent article posits that the failure of recent ‘health in all policies’ is due to health actors being unwilling to engage in the complex systems of other sectors, and, a lack of understanding among government policy makers about the social determinants of health that drive poor health outcomes (Breton, 2016).

It is, I feel, an antiquated limitation, in this age of ‘big data’, and, the ease by which powerful analytic tools can be applied to large datasets using relatively small computing power, that I do not collect such information at the point of care. A recent study by Williams and colleagues demonstrated the reality of the system-level difficulty one encounters to change long-standing practices of not collecting information, or even sharing data across clinical sectors. The researchers, working with health registration clerks at two urban hospitals, attempted to add one
question to the registration process: ‘Do you consider yourself to be of Aboriginal descent?’ As a direct result of the reluctance of data entry clerks to make a change in practice, which arguably would facilitate better information on a variety of health outcomes, this study was prevented from moving beyond the pilot stage (Williams-Roberts, 2017).

I would recommend that future qualitative inquiries delve deeper into specific sub-populations I identified as particularly at risk of high-cost use (multiple chronic conditions, mental health and addictions, lower socio-economic status and palliative care/long-term care residents) to try and understand where cost savings could be achieved. Perhaps emergency room visits could be reduced by embedding geriatricians into long-term facilities themselves, or, perhaps health system workers could be incentivized to connect multi-morbid or mental health clients to a trusted primary care practitioner.

Finally, there is a lack of high-quality intervention research in population and public health. Policy change and programs geared to the ‘problem’ of high-cost health care use have been implemented but rarely comprehensively evaluated. I recommend any interventions aimed at improvement should be accompanied by robust plans to evaluate their impacts. Readers will recall from Chapter 1, Swedish researchers and health policy makers have focused their work on a population of ‘complex, high-need, high-cost elderly’. The work has been largely attributed to the story of ‘Esther’; who, after 5.5 hours in the Swedish medical system had provided her medical history 36 times before she was admitted to hospital and provided treatment. The ‘Esther model’ in Sweden, a program designed to co-ordinate clinical care around being mindful about “what is best for Esther” at each decision, is considered one of the pioneering programs of patient involvement in health system improvement (Vackerberg, Levander, & Thor, 2016). The ‘Esther Model’ could very well be a ‘gold standard’ approach to intervening on a complex, high-
need, high-cost elderly patient; however, there is a lack of research studies aiming to understand the impact of the intervention and its applicability to other contexts.

5.5 Overall conclusions

Readers will recall the two, contradictory, opinions on high-cost user research in the *Annals of Family Medicine* – one arguing that focusing interventions on a narrow high-cost population will add another programmatic silo to already fragmented health care systems (Newton and Lefebvre, 2015). The other opinion argues these populations are exactly the ones that are ripe for intervention, and, a key to health system transformation (Emeche, 2015).

The findings of the present studies are not incongruent with Newton and Lefebvre’s or Emeche’s opinions. By demonstrating quantitatively that socio-demographic factors (whether unstable housing or low income) are associated with high-cost health care use in Saskatchewan, I hope to provide evidence for programmatic interventions for a high-cost, high-need population. I make no claim, one way or the other, as it was not the objective of the current studies, as to what types of programs geared to this population will work to reduce health care costs and improve health outcomes. In reality, these findings should be used in conjunction with current efforts in Saskatchewan focused on primary care system transformation.

Without the ability to quantify cost savings and using data to drive decision-making – policy and decision-makers are left to ideological constructs in developing health system resource planning. Health care economics, health care service use and population health data, both quantitative and
Results from these studies suggest that a point for policy-maker intervention could be by addressing the underlying cause of poverty that is associated with both poorer health outcomes and high-cost health care use. Addressing these underlying causes could be, arguably, more cost effective than improving an individual’s access to health care services (also known as case/care management models of intervention) or management of their specific health issue (disease management intervention models). Addressing the underlying issue of poverty permits a ‘health in all policy’ approach as opposed to silos programmatic interventions aimed at health conditions which leave individuals in the same living conditions that caused their worse health outcomes in the first place (Aue, Roosen, & Jensen, 2016).

Policy and decision-makers require actionable information. Providing descriptive epidemiology on high-cost health care users is not enough – to influence policy and decision-making, actionable evidence must be provided. The methods and analyses conducted to arrive at a reasonable conclusion about what is driving high-cost health care use are often complex. It is incumbent upon implementation scientists to know how to communicate complex findings in a simple, easy to understand manner, in addition to focusing on what is amenable to change – both at the individual or population-level.

Based on experience in conducting knowledge translation efforts from the present work, I suggest that both qualitative and quantitative data be used to illustrate findings. Qualitative data, with good transferability of findings, should be collected in concert with quantitative study objectives. Human beings, as decision-makers are, are compelled by stories. For example, in one
experience of presenting my findings to an audience of over 100 public health professionals across Canada, all of the quantitative data was virtually ignored, but, in both the questions and subsequent evaluations of the presentation, participants remembered and expressed how powerful the story provided by a street outreach nurse was in illustrating the findings. The problem with qualitative data, sometimes, is the lack of transferability when stories relate to a small number of individuals; however, if they are sourced carefully, these stories can add great value to the knowledge translation.

My findings are, I feel, a small piece of the puzzle associated with health system transformation. By providing quantitative evidence on the issue, including costing data, I can move towards action. American studies are likely further ahead given their payment model – there are financial incentives in the for-profit system to find efficiencies; it is interesting that health organizations driven by profit margins for shareholders have come to the conclusion that delivering housing programs and wrap-around care models are cost effective (but, arguably, more importantly, improved health/quality of life for the individual). I have an opportunity in Saskatchewan, the birthplace of universal health care in Canada, to use the quantitative evidence generated here at home and around the world to create a more sustainable model of universal health care – while improving the health of the population and creating a better health care system for all patients and providers.

The practical, conceptual and policy implications of this research are many – there are models that work elsewhere (‘What about Ester?’), American for-profit companies are investing in housing, and, perhaps by bringing together both quantitative and qualitative perspectives I can
galvanize policy makers to address the underlying social determinants of health that are impacting health care costs – but more importantly – human lives lived with dignity.
5.6 References


Breton, E. (2016). A sophisticated architecture is indeed necessary for the implementation of health in all policies but not enough. Comment on: “Understanding the role of public administration in implementing action on the social determinants of health and health


https://doi.org/10.2196/jmir.8508