

UNIVERSITY OF CALGARY

Enhancing Primary Care Electronic Medical Record (EMR) Data in Alberta by Quality
Assessment, Data Processing, and Linkage to Administrative Data

by

Stephanie Garies

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE
DEGREE OF DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN COMMUNITY HEALTH SCIENCES

CALGARY, ALBERTA

JULY, 2020

© Stephanie Garies 2020

ABSTRACT

The growth of electronic medical record (EMR) systems in healthcare settings has created opportunities for EMR data to be reused for secondary purposes. Since EMR data are generated from clinical and administrative processes, the suitability for other uses (e.g. surveillance or research) is questionable. Assessing data quality is important for understanding the database contents, identifying potential limitations or biases, and determining how ‘fit for purpose’ the data are. This thesis focused on evaluating and improving the quality of primary care EMR data in Alberta. Data quality, which is highly contextual, was examined from the perspective of use for hypertension surveillance, as hypertension is a prevalent chronic condition associated with poor health outcomes and high cost implications.

The first part of this thesis involved developing a comprehensive description of EMR data capture, extraction, and processing by the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) in Alberta. The second section presented a data quality assessment using CPCSSN data elements relevant to hypertension surveillance. The third part explored multiple imputation and a pattern-matching algorithm for improving smoking status records in the EMR data. Lastly, EMR and administrative data for a cohort of hypertensive patients were linked and described.

The CPCSSN process documentation and data quality assessment created novel, useful, and comprehensive information for data users. CPCSSN data appear to be suitable for hypertension surveillance, though caution is warranted for several variables of inconsistent quality. Multiple imputation improved completeness of patient smoking statuses, but the lack of an appropriate external reference source made confirming accuracy difficult. The pattern-matching algorithm demonstrated high accuracy for categorizing smoking status; however, it

missed classifying 24% of patients. Lastly, EMR data for 6,307 hypertensive patients were successfully linked to five administrative databases. Although this linked sample is relatively small and may be subject to selection bias (limiting the generalizability for surveillance purposes), the cohort could be useful for health outcomes research or validating elements in the EMR or administrative databases. This work has informed the development of more efficient processes for EMR-administrative linkages. Data quality assessment outcomes will be made available to inform various types of CPCSSN data users.

PREFACE

As part of the program of study outlined in this thesis, the following four manuscripts have been published in or submitted to peer reviewed journals (as described below). For each paper, Stephanie Garies led the study design, analyses, interpretation, manuscript writing, and journal submission. The thesis committee (Drs. Hude Quan, Tyler Williamson, Kerry McBrien, and Neil Drummond) and other co-authors guided the work and contributed important intellectual content and feedback to each manuscript.

Chapter 2 has been published as: Garies S, Cummings M, Forst B, McBrien K, Soos B, Taylor M, Drummond N, Manca D, Duerksen K, Quan H, Williamson T. Achieving quality primary care data: a description of the Canadian Primary Care Sentinel Surveillance Network data capture, extraction and processing in Alberta. *International Journal of Population Data Science* 2019; 4:2.

Chapter 3 was submitted to *BMC Public Health* in November 2019 and is currently in revision as of June 2020: Garies S, McBrien K, Quan H, Manca D, Drummond N, Williamson T. A data quality assessment to inform hypertension surveillance using primary care electronic medical record data from Alberta, Canada.

Chapter 4 has been published as: Garies S, Cummings M, Quan H, McBrien K, Drummond N, Manca D, Williamson T. Methods to improve the quality of smoking records in a primary care EMR database: exploring multiple imputation and pattern-matching algorithms. *BMC Medical Informatics and Decision Making* 2020; 20:54.

Chapter 5 was accepted in *BMJ Health & Care Informatics* on June 29, 2020: Garies S, Youngson E, Soos B, Forst B, Duerksen K, Manca D, McBrien K, Drummond N, Quan H,

Williamson T. Primary care EMR and administrative data linkage in Alberta, Canada: describing the suitability for hypertension surveillance.

This thesis has been professionally edited by Catherine Leipciger for formatting, grammar, and flow (with no modifications made to the content).

ACKNOWLEDGEMENTS

There are many people to thank for supporting me throughout this journey. My supervisors, Dr. Hude Quan and Dr. Tyler Williamson, have been an incredible source of wisdom, encouragement, and mentorship. Their enthusiasm for health research and data is inspiring and has kept me motivated throughout this work! A sincere thank you for all the effort they made to ensure I could succeed in my graduate studies.

A huge thanks to each of my committee members. Dr. Neil Drummond, who started my research career by bringing me on to the CPCSSN team in 2012, has generously provided countless opportunities to advance my professional skills and experiences. Dr. Kerry McBrien has enhanced this thesis with her unique insights as both a family physician and health services researcher; I am thankful for her thoughtful and constructive feedback on this and other work.

I am grateful to collaborate with such a talented and helpful team in Alberta, who have made valuable contributions to this body of work: Dr. Donna Manca, Dr. Michael Cummings, Boglarka Soos, Brian Forst, Kimberley Duerksen, Matt Taylor, and Erik Youngson.

I would also like to recognize the contribution made by the hundreds of sentinel family physicians and nurse practitioners who participate in CPCSSN and especially to those who kindly agreed to permit the linkage of their EMR data for this research.

Lastly, an important thank you to my wonderful husband and son, Heath and Smith – whose unwavering encouragement, support, and patience have truly made this all possible.

I am very appreciative for the funding I received to complete this work, largely through a multi-year Graduate Studentship in Health Innovation from Alberta Innovates and an Alberta SPOR Graduate Studentship, as well as additional graduate training support and travel awards

from various groups at the University of Calgary (Department of Community Health Sciences; Cumming School of Medicine; W21C).

TABLE OF CONTENTS

Abstract.....	ii
Preface.....	iv
Acknowledgements.....	vi
Table of Contents.....	viii
List of Tables.....	xi
List of Figures.....	xii
List of Abbreviations.....	xiii
Epigraph.....	xvi
CHAPTER ONE: AN INTRODUCTION TO PRIMARY CARE ELECTRONIC MEDICAL RECORD (EMR) DATA AND EMR DATA QUALITY.....	1
1.1 Overview.....	2
1.1.1 Overview of the Topic.....	2
1.1.2 Research Objectives.....	2
1.2 Background.....	3
1.2.1 Primary Care EMR Data.....	3
1.2.2 The Canadian Primary Care Sentinel Surveillance Network (CPCSSN).....	5
1.2.3 Challenges of Using EMR Data for Secondary Purposes.....	6
1.2.4 Defining and Assessing EMR Data Quality.....	8
1.2.5 Methods for Improving EMR Data Quality.....	9
1.2.6 Data for Chronic Disease Surveillance.....	10
1.3 Thesis Outline.....	12
CHAPTER TWO: ACHIEVING QUALITY PRIMARY CARE DATA: A DESCRIPTION OF THE CANADIAN PRIMARY CARE SENTINEL SURVEILLANCE NETWORK DATA CAPTURE, EXTRACTION AND PROCESSING IN ALBERTA.....	14
2.1 Abstract.....	15
2.2 Background.....	15
2.3 Approach.....	17
2.3.1 Original Data Source.....	17
2.3.2 Data Stewardship.....	19
2.3.3 Data Extraction.....	19
2.3.4 Data Processing & Data Provenance.....	20
2.3.5 Mapping from Original Values to Standardized Values.....	21
2.3.6 Data Processing Validation.....	22
2.3.7 Audit Trail.....	23
2.4 Discussion.....	23
2.4.1 Lessons Learned.....	24
2.4.2 Ongoing Development / Future Work.....	25
2.4.3 Limitations.....	26
2.5 Conclusion.....	28

CHAPTER THREE: A DATA QUALITY ASSESSMENT TO INFORM HYPERTENSION SURVEILLANCE USING PRIMARY CARE ELECTRONIC MEDICAL RECORD DATA...	37
3.1 Abstract	38
3.2 Background	39
3.3 Methods.....	40
3.3.1 Data Source	40
3.3.2 Patient Sample	42
3.3.3 Data Quality Assessment	42
3.4 Results.....	44
3.4.1 Patient Demographics	44
3.4.2 Height, Weight, BMI	45
3.4.3 Blood Pressure	46
3.4.4 Hypertensive Medications	47
3.4.5 Smoking and Alcohol Status.....	47
3.4.6 External Validity	48
3.5 Discussion.....	48
3.5.1 Limitations	51
3.6 Conclusion	52
 CHAPTER FOUR: METHODS TO IMPROVE THE QUALITY OF SMOKING RECORDS IN A PRIMARY CARE EMR DATABASE: EXPLORING MULTIPLE IMPUTATION AND PATTERN-MATCHING ALGORITHMS.....	63
4.1 Abstract.....	64
4.2 Background.....	65
4.3 Methods.....	67
4.3.1 Data Source	67
4.3.2 Sample.....	67
4.3.3 Defining Smoking Status	68
4.3.4 Defining Demographic and Clinical Variables	69
4.3.5 Methods for Improving Completeness.....	69
4.3.6 Analysis.....	71
4.4 Results.....	72
4.4.1 Methods 1 and 2 (Multiple Imputation).....	74
4.4.2 Method 3 (Pattern-Matching Algorithm).....	74
4.5 Discussion	75
4.5.1 Limitations	78
4.6 Conclusion	79
 CHAPTER FIVE: PRIMARY CARE EMR AND ADMINISTRATIVE DATA LINKAGE IN ALBERTA, CANADA: DESCRIBING THE SUITABILITY FOR HYPERTENSION SURVEILLANCE.....	86
5.1 Abstract.....	87
5.2 Background.....	88
5.3 Methods.....	89
5.3.1 Data Sources	89
5.3.1.1 Primary Care EMR Data.....	89

5.3.1.2 Administrative Data	90
5.3.2 Data Linkage	91
5.3.3 Defining Hypertension	92
5.3.4 Analysis	92
5.3.5 Ethics Approval	93
5.4 Results	93
5.4.1 Data Linkage	93
5.4.2 Patient Cohort Comparisons	94
5.4.3 Provider Comparisons	95
5.4.4 Hypertension Definitions	95
5.5 Discussion	96
5.5.1 Advantages	96
5.5.1.1 Comprehensiveness	96
5.5.1.2 Completeness	97
5.5.1.3 Timeliness	97
5.5.2 Limitation and Potential Biases	98
5.5.2.1 Coverage of Capture	98
5.5.2.2 Completeness & Accuracy	98
5.5.2.3 Misclassification	99
5.6 Conclusion	100
 CHAPTER SIX: SUMMARY	 107
6.1 Summary of Key Findings	108
6.1.1 Describing Data Quality Within a Hypertension Context	108
6.1.2 Post-Extraction Methods for Improving EMR Data Quality	109
6.1.3 Linking EMR Data with Administrative Data	111
6.2 Implications for EMR Data Users	113
6.3 Strengths and Limitations	115
6.3.1 Strengths	115
6.3.2 Limitations	116
6.4 Future Directions	118
6.4.1 Web-Based Data Quality Reporting	118
6.4.2 Additional Data Quality Improvements	119
6.4.3 Future Areas of Research	120
6.5 Conclusion	121
 REFERENCES	 123
 APPENDIX A: SUPPLEMENTARY TABLE 1: COMPARISON OF PATIENT CHARACTERISTICS IN THE PROVINCIAL HEALTHCARE REGISTRY AND THE CPCSSN PRIMARY CARE EMR DATABASE FOR ALBERTA	 139
 APPENDIX B: COPYRIGHT PERMISSIONS	 141

LIST OF TABLES

Table 2.1 Data quality assessment documentation and reporting recommendations	29
Table 3.1 Missingness and summary statistics for patient demographic information	53
Table 3.2 Missingness and summary statistics for physical measurement values.....	54
Table 3.3 Missingness and summary statistics for anti-hypertensive medications	56
Table 3.4 Missingness and summary statistics for risk factor records.....	58
Table 3.5 Prevalence comparisons of adults with hypertension in various data sources.....	59
Table 4.1 Demographic and clinical characteristics of patients with and without smoking status recorded in their EMR.....	81
Table 4.2 Comparison of different improvement methods for smoking categories	82
Table 4.3 Comparison of the pattern-matching algorithm classification to data review	83
Table 5.1 Comparison of characteristics for hypertension patients with and without linked data	101
Table 5.2 Comparison of hypertensive patient characteristics in the CPCSSN database in Alberta and the linked EMR-administrative cohort	102
Table 5.3 Comparison of family physician characteristics in Alberta and in the CPCSSN data	103

LIST OF FIGURES

Figure 2.1 Potential sources of bias and data quality issues in Canadian primary care EMR data	35
Figure 2.2 CPCSSN data pipeline.....	36
Figure 3.1 Summary of completeness of select EMR data elements by EMR type and clinic for patients with hypertension.....	60
Figure 3.2 Paired height and weight measurements in patients with hypertension.....	61
Figure 3.3 Differences in subsequent patient weight measurements throughout time in patients with hypertension.....	62
Figure 4.1 Flow diagram for the pattern-matching smoking status classifier (Method 3).....	84
Figure 4.2 Flow diagram for patients in the cohort study	85
Figure 5.1 Data flow and linkage process for primary care EMR and administrative data in Alberta.....	104
Figure 5.2 Flow diagram of patient selection into the linked hypertension cohort	105
Figure 5.3 Patients in the linked data cohort meeting hypertension case criteria for administrative and EMR definitions by sex and age group (n=6,307)	106
Box 5.1 Criteria for hypertension definitions in administrative and EMR data.....	107
Box 5.2 Criteria for a robust chronic disease surveillance system	108

LIST OF ABBREVIATIONS

AB	Alberta
AHCIP	Alberta Health Care Insurance Plan
AHS	Alberta Health Services
ASP	Application Service Provider
ATC	Anatomical Therapeutic Chemical [classification system]
BMI	Body Mass Index
BP	Blood Pressure
CAC	Centre for Advanced Computing
CCDSS	Canadian Chronic Disease Surveillance System
CCHS	Canadian Community Health Survey
CHMS	Canadian Health Measures Survey
CI	Confidence Interval
CIHI	Canadian Institute for Health Information
CII	Community Information Integration
cm	Centimetres
CPCSSN	Canadian Primary Care Sentinel Surveillance Network
CPRD	Clinical Practice Research Datalink
CSV	Comma Separated Values
CTADS	Canadian Tobacco, Alcohol, and Drugs Survey
DAD	Discharge Abstract Database
DBP	Diastolic Blood Pressure
DIN	Drug Identification Number

eMERGE	Electronic Medical Records and Genomics [Network]
EHR	Electronic Health Record
EMIS	Egton Medical Information Systems
EMR	Electronic Medical Record
EMRALD	Electronic Medical Record Administrative data Linked Database
ERD	Entity Relationship Diagram
FSA	Forward Sortation Area
GIS	Geographic Information System
ICD-9	International Classification of Disease version 9
ICD-10-CA	International Classification of Disease version 10, Canadian modification
ICES	Institute for Clinical Evaluative Sciences
ICPC	International Classification of Primary Care
IQR	Interquartile Range
kg	Kilogram
kg/m ²	Kilograms per Metres Squared
LOINC	Logical Observation Identifiers Names and Codes
MAR	Missing at Random
MCHP	Manitoba Centre for Health Policy
MI	Multiple Imputation
MICE	Multivariate Imputation by Chained Equations
MNAR	Missing Not at Random
mmHg	Millimetres Mercury
NACRS	National Ambulatory Care Reporting System

NAPCRen	Northern Alberta Primary Care Research Network
NIVEL-PCD	Netherlands Institute for Health Services Research Primary Care Database
NLP	Natural Language Processing
NPV	Negative Predictive Value
PBRN	Practice-Based Research Network
PDF	Portable Document Format
PEFR	Peak Expiratory Flow Rate
PHN	Personal Healthcare Number
PIN	Pharmaceutical Information Network
PPV	Positive Predictive Value
SAPCRen	Southern Alberta Primary Care Research Network
SBP	Systolic Blood Pressure
SD	Standard Deviation
SFTP	Secure File Transfer Protocol
SNOMED	Systematized Nomenclature of Medicine
SOAP	Subjective, Objective, Assessment, and Plan
SQL	Structured Query Language
SPOR	Strategies for Patient Oriented Research
STROBE	Strengthening the Reporting of Observational Studies in Epidemiology
SUPPORT	Support for People and Patient-Oriented Research and Trials
UTOPIAN	University of Toronto Practice-based Research Network
UK	United Kingdom
US	United States

EPIGRAPH

Mo' data, mo' problems.

–The Notorious B.I.G. (maybe)

**CHAPTER ONE: AN INTRODUCTION TO PRIMARY CARE
ELECTRONIC MEDICAL RECORD (EMR) DATA AND EMR DATA
QUALITY**

1.1 Overview

1.1.1 Overview of the Topic

As health information becomes increasingly digitized, opportunities to use these data for secondary purposes have emerged, which includes disease surveillance and health research, as well as novel concepts such as precision medicine and learning health systems. The use of electronic medical record (EMR) systems in Canadian healthcare settings has grown substantially over the last decade.^{1,2} Given that the primary functions of EMR systems are for clinical care and administrative tasks, there are concerns about the quality of data when used for secondary purposes. Outcomes derived from the secondary analyses of EMR data may be misinterpreted or biased if the quality of the data is not evaluated and deemed appropriate for the intended use. As such, the evaluation of data quality is an essential first step before using EMR data for secondary purposes.

The thesis research presented here used a framework developed by Kahn et al.,³ which outlines a set of guidelines for reporting on data quality in distributed networks along the entire data lifecycle. Because data quality is contextual, this work focused on the use of primary care EMR data for hypertension surveillance, as this is a condition affecting at least one in five Canadians and is often diagnosed and managed in primary care settings.⁴⁻⁶

1.1.2 Research Objectives

This program of research aimed to undertake a detailed assessment of data quality using primary care EMR data from Alberta and explore different methods for enhancing data quality post-extraction. The specific objectives were:

1. To develop documentation that describes in detail the capture, extraction, and processing of primary care EMR data in Alberta (Chapter 2);
2. To assess the quality of primary care EMR data within a context of hypertension surveillance (Chapter 3);
3. To test methods for enhancing the quality of smoking information in the EMR database (Chapter 4);
4. To link primary care EMR data with administrative data for enhanced uses of hypertension surveillance and research (Chapter 5).

1.2 Background

1.2.1 Primary Care EMR Data

Electronic records were designed as a digital replacement for paper charts that are used in medical settings by physicians and healthcare providers to maintain their patient records, which includes documenting current and past medical diagnoses, medication history, and diagnostic information such as laboratory results.² The use of EMR systems in Canadian healthcare settings have increased substantially over the last decade. Approximately 20% of Canadian physicians were using an EMR system in 2006;⁷ this increased to more than 80% in 2018, as estimated by the Canadian Physician Survey.⁸ In primary care, where the majority of healthcare encounters occur, the vast amount of detailed clinical information captured in EMRs has the potential to inform health research, public health surveillance, and quality improvement. These data have become popular for secondary use for a number of reasons: 1) EMRs contain rich, clinically verified information; 2) the cost and time associated with obtaining the data are much less than primary data collection, such as surveys; 3) EMRs contain large volumes of longitudinal data;

and 4) EMR data represent ‘real-world’ populations, unlike randomized clinical trials, and often with bigger sample sizes.^{4,9,10}

Several countries have been collecting and using primary care or general practitioner EMR data for decades, with notable examples from the United States (US), United Kingdom (UK), the Netherlands, and other European countries.¹¹ For instance, the eMERGE network organizes the linkage of clinical EMR data with genetic information from biorepositories in five medical research sites across the US, enabling genotype/phenotype association studies and contributing to the advancement of personalized medicine.¹² eMERGE was initiated and funded by the National Human Genome Research Institute and operates a central administrative coordinating centre to provide support to each of the contributing sites.¹²

The UK’s Clinical Practice Research Datalink (CPRD) is one of the largest of its kind in the world, with general practice EMR data captured from over 14 million currently registered patients.^{13,14} Anonymized patient-level data are collected on a monthly basis from one of two electronic health systems (CPRD Aurum data from English practices using EMIS [formerly, Egton Medical Information Systems] Web software and CPRD GOLD data from UK practices using Vision software).¹⁵ A subset (58%) of practices participating in the CPRD have also consented to linkages with external data sources from hospitalizations, vital statistics (i.e. mortality), deprivation score measures, and disease registries.¹³ The CPRD is jointly sponsored by the Medicines and Healthcare products Regulatory Agency and the National Institute for Health Research.¹⁴ Research using data from the CPRD has resulted in thousands of publications that have generated evidence for best practice guidelines and has contributed to important improvements in patient safety and care in the UK.¹³

The Netherlands Institute for Health Services Research Primary Care Database (NIVEL-PCD) is a collaboration between the Dutch Ministry of Health, several Dutch universities, the national college of general practitioners, and the national association of general practitioners.¹⁶ The NIVEL-PCD has routinely-collected EHR information for about 1.6 million patients from 500 practices, representing approximately 10% of the total Dutch population.¹¹ Providers who contribute to the database include a wide range of primary health care practitioners, such as GPs, physiotherapists, dieticians, speech therapists, primary care psychologists and others.¹⁶ Researchers cannot access identifiable patient information and the use of the data is determined by steering committees that include representatives from various national associations of health care providers.¹⁶

1.2.2 The Canadian Primary Care Sentinel Surveillance Network (CPCSSN)

CPCSSN's development as a cross-jurisdictional primary care EMR database for Canada began in 2005. It remains the largest and only pan-Canadian source of de-identified EMR data from participating family physicians, nurse practitioners, and community pediatricians.¹⁷ Across the country, the data are extracted from multiple EMR systems, then cleaned, coded, and processed by regional practice-based research networks before being transferred to the national data repository at Queen's University, where further processing and de-identification are performed prior to releasing the data to approved researchers.¹⁷ As of December 2019, 1,574 'sentinel' providers from 257 clinic sites were contributing data for over 1.8 million patients. The CPCSSN database includes a large proportion of information captured in an EMR – patient demographics, diagnoses, prescribed medication, physical examination values (e.g. blood pressure, height, weight), behavioural risk factors, allergies, vaccinations, medical procedures,

referrals, physician billing claims, and laboratory results. It does not include physician narrative text or PDF documents (e.g. scanned referral letters or diagnostic imaging), as these can be difficult to properly de-identify and parse into useable information. In addition to the data elements extracted from the EMR, CPCSSN also provides a set of case definitions to identify patients with certain conditions in the database. Most of these definitions have been validated and demonstrate high validity against a reference standard based on manual chart review or CPCSSN records.^{18,19}

The CPCSSN data are a unique resource in Canada and have been used for a variety of purposes, such as epidemiology and disease surveillance,^{6,20–25} health outcomes research,^{26–29} data tools for primary care clinics,³⁰ and methodological investigations.^{18,31–33}

1.2.3 Challenges of Using EMR Data for Secondary Purposes

Primary care EMR data have many benefits and can be applied to a number of different contexts, but there are concerns about the re-use of EMR data for secondary purposes. Because these data are recorded to support individual patient care and for administrative functions, they may not be produced with the same rigor as research data nor may they be appropriate for all secondary uses.³⁴ The extent to which these data can be used for surveillance or research fundamentally depends on the quality of the database, which can be influenced by many factors at various stages of the data flow.⁹ Data quality can be influenced at an individual provider or clinic level, where differences in data entry training, documentation preferences, and extent of use of the EMR can create huge variations in how and where the information is being recorded. The EMR data itself can be affected by having EMR fields or elements that are difficult to extract from or analyze (e.g. free-text notes or scanned documents), as well as by various

extraction or processing methods. Additionally, complexities arise from combining EMR data from multiple regions across Canada (e.g. CPCSSN) and from different types of EMR systems due to variations in provincial/territorial healthcare systems, health information legislation, and a lack of mandatory content standards for EMR vendors in Canada.

The generalizability of EMR databases can also be of concern, especially when using it for epidemiological and health research purposes. EMR data only include people who have accessed the healthcare system and the data may not be appropriate for studying mild, self-resolving, or short-lived conditions, as these patients may not have sought medical care; this can introduce non-participation bias and limit generalizability to the broader population.³⁵ An investigation into the representativeness of patients in the CPCSSN database found that females and older patients were over-represented in the CPCSSN data as compared to the Canadian census, although this is fairly reflective of primary care populations.³¹ Nonetheless, the large sample size of EMR databases (often hundreds of thousands or millions of patients) can help to decrease the effects of sampling bias and improve generalizability through appropriately adjusting for age and sex. EMR databases also excludes providers who use paper-based charts, though this practice is waning. As compared to respondents of Canada's National Physician Survey, providers in the CPCSSN database were more likely to be female, younger, and practicing in academic settings.³¹

Good data quality is essential for ensuring appropriate inferences are made when using the data for analysis; poor data quality may result in incorrect conclusions that could affect patient care, policy decisions, or applications from research. A foundational step when considering primary care EMR data for any secondary use is to first evaluate the quality of the data and determine whether the data are suitable for the intended purpose.

1.2.4 Defining and Assessing EMR Data Quality

Wang and Strong broadly described the concept of quality, suggesting that data should “be intrinsically good, contextually appropriate for the task, clearly represented and accessible to the data consumer.”³⁶ The specific components used to evaluate data quality can vary, as there is no agreed upon ‘gold standard’ for assessing data quality. For example, a 2013 review identified 27 unique terms that were used to describe dimensions of data quality in electronic health record (EHR) data and noted the challenge of contending with inconsistent terminology across the studies.³⁴ The authors grouped these terms into five general categories: completeness, correctness, concordance, plausibility, and currency.³⁴ Furthermore, the methods used to measure data quality are poorly defined and vary greatly in the literature;^{34,37} these include using a gold standard for comparison, determining whether a data element is present or not, calculating agreement between data elements or sources (e.g. sensitivity or positive predictive value), comparing distributions, checking validity, and conducting log reviews.³⁴

Although a lack of standardized terminology, definitions, and assessment methods may appear problematic, it represents the challenge of tailoring data quality evaluations to match the intended purpose. If each type of data use has certain requirements and levels of importance for specific data elements, then it is unlikely that a singular definition or set of standardized dimensions can be applied to all contexts.³⁸ The framework chosen for my thesis research was developed by Kahn et al., which describes a set of data quality assessment and reporting recommendations for studies that use observational clinical and administrative data for secondary purposes, specifically from distributed data networks (such as CPCSSN).³ The recommendations are unique in that they comprehensively account for the lifecycle of the data

while recognizing that any data quality assessment should be conducted within a specific analytic context.

1.2.5 Methods for Improving EMR Data Quality

With the investment in health information systems, technology, and data sciences by Canadian provinces and territories, a continuous cycle of data quality assessment, reporting, and improvement should be embedded into all aspects of EMR data generation and use. High quality data are fundamental for secondary uses (i.e. research, disease surveillance, policy decision-making, clinical practice improvement) and, therefore, strategies to improve deficiencies in the quality of data are needed. These approaches can occur *pre*-data extraction or *post*-data extraction, in addition to more broad policy-based solutions.

- Improving data quality pre-extraction is the most ideal strategy; however, these tend to be more complicated to enact. Examples include addressing data entry behaviours of clinicians or clinical staff through competency-based training and education,³⁹ audit and feedback,⁴⁰⁻⁴² or using data entry clerks for quality checks and to update information in the EMR.⁴³
- Improving data quality post-extraction tends to include computerized/statistical methods or analytical solutions, which are easier to implement than pre-extraction solutions but do not address underlying causes of poor data quality. For instance, the development and validation of definitions for disease identification using multiple criteria helps to mitigate incomplete or inaccurate information in the EMR;^{18,44} the use of advanced techniques like natural language processing to create useable information from unstructured narrative text in the EMR or multiple imputation for

dealing with missing data;⁴⁵ creating cleaning and coding algorithms that map unstructured information to standardized classification systems or terminologies;^{17,46,47} or employing association rule mining, which can be used to infer the presence of health conditions in the absence of complete and accurate patient problem lists by using associations with other data elements (e.g. medication, diagnoses, laboratory results).⁴⁸ The linkage of EMR data to other sources (e.g. administrative data) allows for verification of case definitions and select variables or diagnostic codes.⁴⁹⁻⁵¹

- A third overarching approach addresses systemic deficiencies through policy. This includes mandating content standards for EMR vendors;⁵² building EMR software to include more structured data fields and a better-designed interface to improve data entry efficiencies;⁵³ or providing financial incentives to attain minimum levels of data quality and promote ‘meaningful EMR use’.^{13,35,54,55} Large-scale initiatives such as these pose implementation challenges, as well as the potential for major negative effects.

Ideally, a variety of multifaceted, ongoing approaches for improving EMR data quality should be considered, though this is largely dependent on available resources, clinician time, the development of an EMR data science workforce, and political planning.

1.2.6 Data for Chronic Disease Surveillance

In Canada, chronic disease surveillance is often conducted using administrative data or surveys. For example, the Canadian Chronic Disease Surveillance System (CCDSS) is comprised of administrative data from each province and territory, which includes the health

insurance registry database linked to physician billing claims, hospitalizations, and (where available) prescription drug databases.⁵⁶ Administrative data have the advantage of being population-based with data captured from nearly all healthcare encounters within a province or territory. However, administrative data lack important patient-level details about physical measurements (e.g. blood pressure, height, weight), behavioural risk factors (e.g. smoking, alcohol use, diet, exercise), and aspects of primary care activity outside of billing claims. In contrast, national surveys are designed to obtain comprehensive information about the health and healthcare utilization of a representative sample of Canadians, as is achieved through the Canadian Community Health Survey (CCHS).⁵⁷ The limitations with health surveys are the cross-sectional design, which does not allow for following participants over time, and the significant cost of primary data collection repeated every two years.

As primary care EMR data have become more established in Canada, there is an opportunity to use these data to augment our existing surveillance sources. Primary care EMR data captured by CPCSSN nationally represent a relatively small sample of citizens and their healthcare providers (approximately 5% of the 2019 Canadian population of 37.8 million people⁵⁸), but the data are longitudinal, comprehensive, and collected objectively by family physicians, nurses, and clinical staff. National CPCSSN data have been used previously for chronic disease surveillance^{6,20-23} and several provinces have expanded beyond this to utilize primary care EMR data linked to administrative databases. For instance, the Manitoba Centre for Health Policy (MCHP), housed within the University of Manitoba, collects and links a large variety of provincial data such as health (including primary care EMR data), social, justice, education, and population-level registries.^{59,60} In Ontario, the Institute for Clinical Evaluative Sciences (ICES)⁶¹ has established processes for linking primary care EMR data from two groups

(the EMR and Administrative data Linked Database / ‘EMRALD’;⁵¹ University of Toronto Practice-Based Research Network / ‘UTOPIAN’⁶²) with their administrative data holdings.

Linking EMR and administrative data produces a new source that can be used for novel applications, such as enhanced monitoring of patients throughout the entire healthcare system, the addition of valuable population health risk factor data for surveillance (such as smoking or weight), and improved identification of high-risk or high-cost individuals to develop effective strategies for management in the community.^{9,63} However, there are various technical and policy challenges that make EMR-administrative data linkage complex; this work can only be conducted within each province or territory due to specific health legislation, data availability, and data access. In Alberta, primary care EMR-administrative data linkage has not been conducted routinely and has only been attempted on very few occasions.

1.3 Thesis Outline

This manuscript-based thesis focuses on examining the quality of primary care EMR data from Alberta and methods to improve areas of poor data quality. The decision to focus only on data from one province (Alberta) was taken due to the regional differences that exist in CPCSSN data capture, extraction and processing, which necessitated the need for regionally-based data documentation and quality assessment. In addition, data linkage can only be achieved within each province or territory according to the respective health information legislation.

The four chapters that follow have been formatted for publication in peer-reviewed journals. Chapter 2 outlines a comprehensive description of the CPCSSN data lifecycle in Alberta using a framework for data quality reporting in distributed networks.³ This foundational information is critical for users of primary care EMR data to fully understand how the CPCSSN

data are derived, extracted, and processed and has not been previously available in this level of detail until now. Chapter 3 summarizes the quality of EMR data within the context of hypertension; this paper aims to provide a descriptive summary of the data to help potential data users assess the suitability of these data for hypertension surveillance or research. Using these findings, patient smoking status was identified as a data element with poorer quality due to missingness and potential inaccuracies. Chapter 4 examines two methods for improving the quality of smoking records in the EMR database using multiple imputation and a pattern-matching algorithm to categorize patient smoking status for those with hypertension. Chapter 5 describes a third method for improving data quality and enhancing hypertension surveillance through linkages to administrative databases and the impact of data linkage as it relates to chronic disease surveillance. Here, primary care EMR data from Alberta were linked to five administrative sources: Alberta Health Care Insurance Plan (AHCIP) Registry, Discharge Abstract Database (DAD), National Ambulatory Care Reporting System (NACRS), Pharmaceutical Information Network (PIN), and practitioner claims. Lastly, Chapter 6 summarizes the overall quality of primary care EMR data in Alberta within the context of hypertension and what methods are recommended to improve the quality of EMR data, including the strengths and limitations of this work. The implications for various EMR data users (e.g. clinicians, researchers, decision-makers) will be discussed and future directions for this work will be highlighted.

**CHAPTER TWO: ACHIEVING QUALITY PRIMARY CARE DATA: A
DESCRIPTION OF THE CANADIAN PRIMARY CARE SENTINEL
SURVEILLANCE NETWORK DATA CAPTURE, EXTRACTION AND
PROCESSING IN ALBERTA**

This chapter is published as:

Garies S, Cummings M, Forst B, McBrien K, Soos B, Taylor M, Drummond N, Manca D, Duerksen K, Quan H, Williamson T. Achieving quality primary care data: a description of the Canadian Primary Care Sentinel Surveillance Network data capture, extraction and processing in Alberta. *International Journal of Population Data Science* 2019; 4:2.

2.1 Abstract

Introduction: Electronic medical record (EMR) databases have become increasingly popular for secondary purposes, such as health research. The Canadian Primary Care Sentinel Surveillance Network (CPCSSN) is the first and only pan-Canadian primary care EMR data repository, with de-identified health information for almost two million Canadians. Comprehensive and freely available documentation describing the data ‘lifecycle’ is important for assessing potential data quality issues and appropriate interpretation of research findings. Here, we describe the flow and transformation of CPCSSN data in the province of Alberta.

Approach: In Alberta, the data originated from 54 publicly-funded primary care settings, including one community pediatric clinic, with 318 providers contributing de-identified EMR data for 410,951 patients. Data extraction methods have been developed for five different EMR systems, and include both backend and automated frontend extractions. The raw EMR data are transformed according to specific rules, including trimming implausible values, converting values and free text to standard terminologies or classification systems, and structuring the data into a common CPCSSN format. Following local data extraction and processing, the data are transferred to a central repository and made available for research and disease surveillance.

Conclusion: This paper aims to provide important contextual information to future CPCSSN data users.

2.2 Background

The proportion of family physicians using electronic medical record (EMR) systems has been steadily increasing in Canada, with an estimated 85% reporting some degree of EMR use in practice in 2017.¹ Coupled with advancements in computing power, data storage capacity, and

analytic methods for large datasets, primary care EMRs are becoming a popular source of data for health research and other secondary uses.^{9,64} The Canadian Primary Care Sentinel Surveillance Network (CPCSSN) maintains a pan-Canadian primary care EMR database made available for health research, disease surveillance, and clinical quality improvement.^{9,17} The CPCSSN data have been used to develop and validate case definitions for many conditions relevant to primary care,^{18,65,66} report on the epidemiology of chronic conditions across Canada,^{6,20–25,67} contribute to EMR data methodology and quality,^{31,32,43,68} and conduct various health outcomes research.^{26,28,69–71} However, in order to ensure the CPCSSN data can be used for these purposes, sophisticated cleaning and processing techniques are required to transform the raw EMR data extracted from primary care settings into useable information. This includes coding free text, trimming implausible values, and standardizing the data into a common format. The complexity of these tasks becomes multiplied due to variation in data quality and quantity between providers, clinics, provinces/territories, and EMR systems.

Comprehensive documentation describing the data flow is a critical part of ensuring high data quality; more specifically, understanding where the data originate, the process for extraction and transformation, and how data samples are created for analysis. Without this information, it would be challenging to determine whether the data are suitable for the intended use and any research findings would risk misinterpretation. The Kahn et al. framework for reporting data quality in distributed networks provides a template for thoroughly describing the entire lifecycle of data, while aligning with several STROBE (‘Strengthening the Reporting of *Observational Studies in Epidemiology*) recommendations.³ Although an overview of the CPCSSN organization and database has been previously published,¹⁷ a detailed description of the data capture and processing outlined in the Kahn framework have not, to date, been available for the

CPCSSN database. The purpose of this paper is to provide a technical report of the generation, extraction, and processing of primary care EMR data in Alberta. We anticipate that this work will support prospective users in their understanding of the CPCSSN data context and quality.

2.3 Approach

2.3.1 Original Data Source

Canada has a universal, publicly funded healthcare system with each province or territory managing the delivery of healthcare according to their own jurisdiction's needs. Because of this, variation exists between each province and territory, in terms of health information legislation, compensation models for physicians, medications available for prescribing, and healthcare organization. Primary care is considered the first point of contact for patients in the health system and consists of medical care, preventive services, chronic disease management, and referral to specialist physicians. In Alberta, primary care providers often work in multi-disciplinary team practices that include family physicians and other allied health professionals such as nurses, nurse practitioners, dietitians, mental health specialists, and many others. Patients are free to visit any primary care practice of their choosing at any time, although attempts have been made to formally attach people to a singular primary care provider to promote better continuity of care. During a primary care consultation, the patient's relevant information is entered into the EMR system, usually manually, though exactly what is recorded depends on provider preferences and EMR system input options. Generally, information such as demographics, diagnoses (or symptoms or suspected conditions), prescribed medications, physical measurements (such as blood pressure, height, weight), risk factors (smoking, alcohol use), allergies, vaccinations, and medical procedures are entered by the clinicians or by clinic staff (e.g. physician, nurse,

receptionist, clerk). Results from laboratory service providers are electronically merged into the patient record.

In 2008, the Public Health Agency of Canada funded the initial development of CPCSSN, with the goal of creating a pan-Canadian repository of primary care EMR data that could be used for surveillance and research. At the present time, CPCSSN has grown to over 1,200 primary care providers from eight provinces and territories who contribute de-identified data for close to two million patients.⁷² CPCSSN is a collaboration of eleven regional primary care practice-based research networks (PBRN) across Canada. Primary care PBRNs are typically multidisciplinary collaborations between family physicians, allied health care providers, and university-affiliated primary care researchers with the purpose of addressing questions deriving from clinical practice and facilitating improvement in primary care settings. Each PBRN recruits ‘sentinel’ primary care providers (family physicians, nurse practitioners) and community pediatricians to the project and is also responsible for local extraction, cleaning, and processing of the EMR data prior to transferring the regional data to the central repository located at Queen’s University in Kingston, Ontario.¹⁷ Additional details about the organization and data processes of the national CPCSSN database have been described elsewhere.^{17,72} Across the country, CPCSSN actively extracts data from eleven different EMR systems, which vary by region; however, this paper is focused on EMR data obtained by the two regional PBRNs in the province of Alberta – the Northern and Southern Primary Care Research Networks (NAPCRen and SAPCRen, respectively). Between both Alberta networks, 318 providers from 53 primary care practices and one community pediatric clinic contribute de-identified data from 410,951 patients in the province, as of December 31, 2018 (Table 2.1). This represents just over 5% of the total population of family physicians in Alberta.⁷³

2.3.2 Data Stewardship

NAPCReN and SAPCReN are embedded within the Departments of Family Medicine at the University of Alberta (NAPCReN) and the University of Calgary (SAPCReN), and are funded through various project grants and departmental contributions. As outlined in Alberta's health legislation, health care service providers, such as nurses and physicians, may be designated as custodians of the patient data. Both NAPCReN and SAPCReN have research ethics board approval for the CPCSSN project at their respective institutions. Through research agreements with each data custodial 'sentinel' provider, NAPCReN and SAPCReN are the data stewards who are permitted to extract and process de-identified EMR data before transferring to the central CPCSSN repository. Sentinels contribute to CPCSSN voluntarily, without financial compensation, and for their participation receive their panel data in the form of routine feedback reports and an online query tool, as well as membership in their respective PBRNs. Patients are informed of the CPCSSN project through information posters and brochures in each clinic, and are able to opt-out of the database at any time by contacting a PBRN staff member or informing their provider. If this does occur, the CPCSSN data manager will obtain a unique identifier associated with the patient's record (EMR ID) from the clinic and add an exclusion for this ID to the extraction code.

2.3.3 Data Extraction

Table 2.1 provides details about the data extraction and processing. Alberta has one of the highest rates of EMR use in family practice of Canadian provinces and territories at over 86%.¹ NAPCReN and SAPCReN have extraction processes for five EMR systems in Alberta: Wolf, Med Access, Practice Solutions Suite, Accuro, and Healthquest. Data extraction occurs every six

months. The extraction process depends on the EMR system and whether a hosted/Application Service Provider (ASP) or a local/non-ASP system is used. Two systems require backend extractions through the vendor (Wolf, Practice Solutions Suite), two allow backend extraction by a CPCSSN data manager (Accuro, Healthquest), and one is an automated frontend extraction (Med Access). A full list of data elements extracted is available in the CPCSSN Data Dictionary.⁷⁴

An EMR-to-CPCSSN patient ID mapping file is also created at the time of extraction. This file contains the original EMR patient ID that is assigned to each patient within a clinic EMR system, the CPCSSN network ID and CPCSSN site ID for the clinic from which the data are extracted, and the unique CPCSSN patient ID that is randomly assigned to each patient. The file is stored separately from the de-identified health data and is used to ensure that, during data extraction, CPCSSN patient IDs are assigned uniquely and consistently to each EMR patient ID, as well as to facilitate re-identification for practice quality improvement or data linkage.

2.3.4 Data Processing & Data Provenance

CPCSSN has developed a number of processing steps that transform raw EMR data into a format that is compatible with the CPCSSN database structure. Many elements in the CPCSSN database contain both an ‘original’ field and a CPCSSN ‘calculated’ or ‘coded’ field. At the regional level, specific processing work is undertaken following data extraction, including general data cleaning for specific fields (e.g. standardizing dates, deleting empty or duplicate fields), trimming values to fit allowable ranges, mapping original data to standard classification or coding systems, and some additional transformations to create new variables (Table 2.1). Each CPCSSN-affiliated PBRN may have different processes for the transformation of their local data;

due to shared data manager resources between NAPCReN and SAPCReN, the processes described here are the same for both Alberta networks unless otherwise specified in Table 2.1.

2.3.5 Mapping from Original Values to Standardized Values

Standardization of numerical CPCSSN data is relatively straightforward. For physical examination or laboratory data, all values are converted to one common set of units. Bounds are also placed on the data, and values that lie beyond these ranges are not coded into the database. Age and date ranges are also restricted. Any age or year entry (e.g. patient birth year) must indicate that the patient is less than or equal to 120 years old at the time of data extraction. Dates are constrained by type: those relating to EMR record creation date must fall within the range of 1990-01-01 to the extraction date, while condition onset or procedure performed dates must be between the extraction date minus 120 years and the extraction date.

The original EMR data are mapped to standard classification or coding systems, such as Anatomical Therapeutic Chemical (ATC) codes for medication, International Classification of Disease version nine (ICD-9) codes for diagnostic text, and Logical Observation Identifiers Names and Codes (LOINC) for laboratory values. Unstructured risk factor information (such as smoking and alcohol use) is categorized into risk type (and additional categories, if the data are present) using a pattern matching approach. Some additional transformations are performed to create new variables, including BMI calculated from weight and height (if no BMI exists), and disease cases indexed according to CPCSSN's validated definitions.¹⁸ Material and social deprivation quintiles based on mapping postal codes to the Pampalon Index⁷⁵ will be integrated into the database in the near future.

2.3.6 Data Processing Validation

To validate the transformation processes, both manual and automated routines are applied to the CPCSSN-produced data. Consistency and completeness of extracted data are analyzed through manual file size and number checks. Automated software-based analysis can also be applied, including calculation of per provider statistics (number per clinic, number of patients, number of records) and comparison to values from the previous extraction. Transformation completeness is primarily assessed through review of the verbose log files. These files contain initial and final row counts for each table in the database, and the numbers of rows affected by each process applied to the data. Data generated following the coding/cleaning processes are also assessed by manual viewing of log files; however, the underlying software first undergoes two other tests. To determine how updates to specific coding and/or cleaning algorithms will affect data processing, the software is run against algorithm-specific CPCSSN reference data standards. For example, if changes are made to the exam cleaning algorithm, the software is run on a copy of the exam reference data set. The resulting cleaned data are then compared to the original exam reference set and any changes are assessed. Additionally, an end-to-end test is applied, where all steps of the coding/cleaning software are run on the gold standard data set that was used to define the original eight CPCSSN disease case definitions.¹⁸ The sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) for these eight diseases are then calculated and compared to values produced by the previous version of the software.

Once a database has been coded, cleaned, and all case detection algorithms run, personal identifiers that are imbedded in free-text are redacted. This is accomplished through matching to eight classes of regular expression patterns, including patient first and last names, provincial health numbers, and phone numbers. For a list of redacted terms, see Table 2.1.

2.3.7 Audit Trail

Verbose logging is produced at each of the three processing stages (extraction, transformation, and coding/cleaning), which serves as an audit trail. One extraction and one transformation log are produced per clinic. Following the transformation stage, databases for all clinics within a network are generally merged, and so there is typically only a single coding/cleaning log for each network. Each line of the log files begins with a timestamp, containing the current date and time, and a message that may contain processing software version information, source or destination file path or numbers of data elements affected by a process. Warning and error messages, including information such as missing source files, processes that failed, or data elements that are beyond permitted ranges, are also recorded.

Each PBRN typically completes data processing on their own secure server that is housed at the Centre for Advanced Computing (CAC) at Queen's University. Once the CPCSSN data are processed and standardized, the data are transferred within the secure network at the CAC to the central CPCSSN repository through simple file transfer. Data from all networks are then combined within the central repository to form the national CPCSSN database.

2.4 Discussion

We have described the process used by NAPCRen and SAPCRen for data extraction, transformation, cleaning/coding, and merging, with the aim of providing important context for current and future users of the CPCSSN data. In the full lifecycle of EMR data, there are many potential opportunities for data quality issues and biases to occur.⁷⁶ Particularly in distributed data networks, variability may occur due to differences in provincial/territorial policies, payment models, delivery of care, availability of EMR systems, and healthcare utilization, among others

(Figure 2.1). Additionally, the local extraction and processing for CPCSSN can vary by regional PBRN, which underscores the importance of having detailed documentation available for data users. This work offers a set of standards that may be used by other PBRNs to describe their own EMR data practices and/or to adopt the procedures described here.

2.4.1 Lessons Learned

The CPCSSN data extraction and transformation processes have been achieved through more than a decade of dedicated work by a national team of data managers. Significant advantages of the CPCSSN data (compared to the source EMR data) include implementation of many automated processes for cleaning raw data and mapping of values or text to standard classification systems (such as the ATC Classification System, LOINC, and ICD-9). Additionally, the development and implementation of 15 (and counting) validated case definitions allow clinicians and researchers to more accurately identify patients with certain conditions. Informal searching through poorly coded or unstructured raw data limits the ability of clinicians to undertake quality improvement projects or introduces biases when conducting research or surveillance. The CPCSSN team places significant importance on having valid case definitions for the CPCSSN database, as demonstrated by an internal ‘Case Definition Working Group’ and its efforts to continually evaluate and improve existing definitions, while working towards the development of new ones.

Collaboration between regional networks and the central CPCSSN data management office at Queen’s University has been essential for the coordination of CPCSSN data-focused tasks, though this is not always easy to achieve. In particular, sound software development principles are vital in developing and coordinating robust extraction, transformation, and coding/cleaning tools between eleven pan-Canadian networks. Use of a version control system,

such as GIT or Mercurial, and tagging of software versions is essential, so that it is clear precisely which version of a given program, plus any associated configuration files, were used to generate a particular data set. Increasing the level of automation to reduce the amount of direct human intervention in the processing pipeline has greatly increased the level of data consistency and reproducibility not only between clinics, but also between extraction cycles. EMR data are surprisingly messy, so software developed to process it quickly becomes complex. Since even carefully written code will contain bugs, by developing test data sets and testing software, coding errors are more easily caught and corrected. Finally, while machine learning methods can be extremely powerful tools for data cleaning and coding, they are not always the best solution. In many cases, simple pattern matches can be implemented more quickly and with comparable or higher accuracy, and can lead to more maintainable software.

A sample of the national CPCSSN database is available to approved university-affiliated researchers after submission of a letter of intent for review by the CPCSSN research committee, research ethics approval at their host university, and payment of data access fees; more details are provided on the CPCSSN website (www.cpcssn.ca). Access to regional CPCSSN data may be permitted by the director of the PBRN upon discussion, research ethics approval (for research studies) and adherence to any network-specific requirements.

2.4.2 Ongoing Development / Future Work

The extraction and transformation processes described here will undoubtedly evolve as CPCSSN continues to recruit primary care providers and expand its data holdings. Future plans include improvements to the risk factor classifier algorithms (e.g. more accurately assigning patients to the correct smoking status), additional coding for diagnoses entered as free text, and

exploration of extraction methods for new EMR systems. Machine learning-based improvements to lab, medication, and referral coding are also underway.

Linkage of the CPCSSN EMR data to administrative data sources (e.g. hospitalizations, emergency department visits, pharmacy dispensed medications) is a major step towards a more complete understanding of a patient's journey through the health care system. Although the CPCSSN data are de-identified, a mapping file generated from each clinic EMR system (containing EMR ID, patient Personal Health Number (PHN), date of birth and sex) can facilitate linkage based on PHN with administrative data held in the provincial health authority's data warehouse. A second mapping file from CPCSSN (which includes patient EMR ID and CPCSSN patient ID) can be used to map the de-identified CPCSSN data to the administrative data using EMR ID. Each regional PBRN may have a different process for data linkage (if at all), largely due to differences in the privacy laws and health information acts within each province/territory. For those PBRNs that conduct data linkage, the CPCSSN EMR data are linked and held within each province/territory and are not part of the national CPCSSN database.

2.4.3 Limitations

This paper describes CPCSSN data from two regional PBRNs in Alberta that collect de-identified EMR data from a sample of primary care providers, which may not reflect the characteristics or practice of all providers in Alberta or of the 37,000+ family physicians⁷³ in other provinces/territories. Although citizens who are registered with the health care system in each province/territory are assigned a unique health identifier (termed Personal Health Number (PHN) in Alberta), identifiable information (including PHN) are not collected as part of the CPCSSN database. This means that CPCSSN currently has no way to identify patients who

receive care at multiple clinics and thus, may be assigned multiple CPCSSN IDs. While progress on rostering patients within Alberta to a specific primary care provider may help to solve this problem, the extent of the issue in the CPCSSN data for Alberta is not known.

Primary care EMR data in Canada are still a relatively new source of health information and there are few existing standards for EMR data entry, extraction, and processing. Thus, some of the processing steps were developed in response to a need at a particular time, rather than based on an established reference standard. Any process that is manually-driven may be prone to error and lessens the ability to replicate work. In addition, this paper is limited to the description of extraction and processes in Alberta, which may vary from the other CPCSSN-contributing PBRNs across Canada. Given the distributed nature of the national CPCSSN project and the scarcity of process documentation available to data users, it is hoped that this work serves as a template for PBRNs to record their data capture and transformation processes. Lastly, both recruitment of CPCSSN primary care providers in Alberta and updates to the technical cleaning and coding work is ongoing, meaning that the information provided here may change over time as providers enter or exit the sample and data processing improvements are implemented. EMR data quality is likely to vary based on data capture; this is related to data entry practices in each clinic by different providers and how well the EMR interface facilitates the recording of information. At present, we do not capture any details about variations in clinical data entry and design of EMR systems. Furthermore, NAPCReN and SAPCReN do not extract healthcare provider notes (e.g. ‘Subjective, Objective, Assessment, and Plan [SOAP] notes) or PDF documents, such as referral letters or diagnostic imaging.

2.5 Conclusion

Primary care EMR data are a valuable source for surveillance and health research, in addition to facilitating more rapid clinical quality improvement and evaluation. There are many complex steps needed to extract and transform raw EMR data into useable information, which have been described here. The comprehensive documentation of the CPCSSN processes in Alberta will enable better understanding and use of these nuanced data, as well as provide a systematic methodology that can be used by other groups working in this domain.

Table 2.1 Data quality assessment documentation and reporting recommendations (as outlined by Kahn et al.³)

Data Capture		
Original Data Source		
Data origin	<p>NAPCRen: EMR data from primary care clinical practices in northern Alberta using one of Accuro (<i>n</i>=1), Healthquest (<i>n</i>=1), Med Access (<i>n</i>=11), or Wolf (<i>n</i>=6) EMR system.</p> <p>As of December 31, 2018, data was captured from: 19 primary care clinics 79 providers (family physicians) 104,622 patients</p>	<p>SAPCRen: EMR data from primary care clinical practices in southern Alberta using one of Med Access (<i>n</i>=12), Practice Solutions (<i>n</i>=1), or Wolf (<i>n</i>=22) EMR system.</p> <p>As of December 31, 2018, data was captured from: 34 primary care clinics + 1 community pediatric clinic 239 providers (family physicians, nurse practitioners, community pediatricians) 306,329 patients</p>
Data capture method	<p>Direct entry by clinicians/clinical staff into EMR system; interface and available templates/data fields may vary by EMR system.</p> <p>Automated upload of laboratory results into EMR from community lab provider.</p>	
Original collection purpose	<p>Patient clinical care and administrative tasks (i.e. billing) in community-based primary care clinics in Alberta.</p>	
Data Steward Information		
Data steward	<p>NAPCRen: Northern Alberta Primary Care Research Network (NAPCRen), one of 11 regional practice-based research networks hosting the CPCSSN project. Housed within the Department of Family Medicine, University of Alberta; Edmonton, Alberta, Canada.</p>	<p>SAPCRen: Southern Alberta Primary Care Research Network (SAPCRen), one of 11 regional practice-based research networks hosting the CPCSSN project. Housed within the Department of Family Medicine, University of Calgary; Calgary, Alberta, Canada.</p>

Database model/structure	<p>The CPCSSN data are stored as a SQL database (either SQL Server or SQLite format) and currently contains 26 tables:</p> <ul style="list-style-type: none"> ▪ Eight tables house clinic, health care provider, or administrative information (e.g. patient-provider assignment, non-identifiable clinic information, database version information) ▪ Fifteen tables hold both original EMR data and processed (cleaned and coded) data ▪ Two tables (DiseaseCase, DiseaseCaseIndicator) contain a list of patients who have been indexed with one or more of the conditions for which CPCSSN has a validated definition; these definitions use data derived from the 15 data tables ▪ The Deprivation table is empty, but will be populated in the future. <p>Each table is assigned an integer primary key labelled: <table_name>_ID. Most data tables also contain the columns:</p> <ul style="list-style-type: none"> ▪ Site_ID (an integer assigned to each primary care practice) and Network_ID (an integer assigned to each PBRN) together form a composite primary key in the Site table, and a composite foreign key in most CPCSSN tables ▪ Patient_ID (randomly assigned integer for each patient) is the primary key in Patient table and a foreign key in most other tables ▪ Encounter_ID is the integer primary key for the Encounter table, but a foreign key in several other tables ▪ Provider_ID is the integer primary key for the Provider table and a foreign key for several tables ▪ All columns are strictly typed, enforced either by the database schema (SQL Server) or external software (SQLite) <p>The Entity Relationship Diagram (ERD), which outlines relationships between all CPCSSN tables, can be found on the CPCSSN website: http://cpcssn.ca/research-resources/cpcssn-data-dictionary-and-erd/</p> <p>For SAPCReN and NAPCReN, extraction, transformation, coding, and cleaning algorithms are programmed in Python.</p>
Data dictionary	The most recent data dictionary is posted on the CPCSSN website: http://cpcssn.ca/research-resources/cpcssn-data-dictionary-and-erd/
Data Processing / Data Provenance	
Data extraction specifications	<p>The general procedure for EMR data extraction in both SAPCReN and NAPCReN is as follows:</p> <ul style="list-style-type: none"> ▪ Access the EMR frontend or backend, either through the vendor or by a CPCSSN data manager <ul style="list-style-type: none"> ○ Wolf & PS Suite: backend extraction completed by Telus Health ○ Accuro & Healthquest: backend extraction conducted by a CPCSSN data manager using a log-in account for both clinic and EMR database ○ Med Access: frontend extraction by a CPCSSN data manager using a log-in account created by the clinic ▪ Select data tables and columns of interest, then: <ul style="list-style-type: none"> ○ Null any columns with identifiable information (e.g. names, addresses, health care numbers) ○ Select patients with an assigned CPCSSN provider or for those who have not been assigned any provider, use the four-cut method⁷⁷ to assign a provider ○ An EMR mapping file is generated which consists of four data items: EMR patient ID (number assigned by each clinic EMR system), Network ID (number that uniquely identifies each CPCSSN regional network), Site ID (number that

	<p>uniquely identifies each clinic within a network) and CPCSSN patient ID (unique patient number randomly assigned by CPCSSN). This file is used to ensure CPCSSN patient IDs are assigned uniquely to each EMR patient ID within each clinic and is held separately from the patient health data. The file can also be used to assist with re-identification for practice quality improvement or linkage with other data sources</p> <ul style="list-style-type: none"> ▪ Transfer data to CPCSSN computers via a secure transfer method (e.g. SFTP) <p><i>See Figure 2.2 for CPCSSN data pipeline.</i></p>
<p>Mappings from original values to standardized values</p>	<p>The CPCSSN database includes both original data extracted from the EMR, and cleaned or coded information resulting from CPCSSN's algorithms (see CPCSSN Data Dictionary⁷⁴). All original data are extracted from the EMR unchanged, except for a small number of EMR-specific code conversions and some parsing.</p> <p><u>General cleaning steps</u></p> <ul style="list-style-type: none"> ▪ Null empty strings and remove leading or trailing whitespace from all records ▪ Normalize postal codes to format A1A 1A1 <ul style="list-style-type: none"> ○ NAPCReN: truncate to forward sortation area code (A1A) ▪ Convert all dates to format yyyy-mm-dd ▪ Verify code formats for ATC, ICD-9, and DIN classification systems ▪ Delete invalid labs where <ul style="list-style-type: none"> ○ lab name is a purely numeric value, or matches a text patterns that indicates it is not a useful lab name (e.g. 'unit', 'patient name', 'date') ○ lab value non-numeric; exceptions are lab values starting with a comparison operator and Hepatitis C ▪ Convert medication dispensed forms and duration units to CPCSSN standard forms ▪ Delete empty or duplicate records ▪ Remove orphaned records (e.g. records with no matching Patient_ID in the Patient table) ▪ Remove patients who only have demographic information and no other data ▪ Parse billing records that contain multiple conditions into individual records <p><u>Allowable ranges of values for NAPCReN and SAPCReN:</u></p> <ul style="list-style-type: none"> ▪ SBP: 50 to 300 mmHg ▪ DBP: 20 to 200 mmHg ▪ Weight: 1 kg to 500 kg ▪ Height: 30 cm to 230 cm ▪ BMI (adult): 5 to 200 kg/m² ▪ Waist circumference: 10 to 300 cm ▪ Waist-to-hip ratio: 0.1 to 1 ▪ PEFr: 0-200 L/min

	<ul style="list-style-type: none"> ▪ Ranges for laboratory values are specific to each of the 44 lab results processed by CPCSSN ▪ Age at onset: ≤120 ▪ Birth year: extraction year (e.g. 2018) minus 120 years ▪ Record creation date: earliest EMR date (1990-01-01) to extraction end date (2018-12-31) ▪ Condition onset or procedure performed date: 1898-12-31 (extraction end date minus 120 years) to 2018-12-31 <p><u>Coding to reference system</u></p> <ul style="list-style-type: none"> ▪ Medication: existing code, DIN, or free text mapped to codes in the ATC Classification System ▪ Laboratory values: existing code or free text mapped to LOINC; free-text mapping is completed for 44 labs ▪ Diagnostic codes: ICD-9 or other codes (e.g. ICD-10, ICPC) found in Health Condition/Profile, Encounter Diagnosis, and Billing tables mapped to ICD-9 codes/classification heading ▪ Diagnostic text: Pattern matches are used to map free text to ICD-9 codes; this mapping can be completed for 48 conditions in the Health Condition/Profile, Encounter Diagnosis, Billing tables ▪ Referrals: original text in Referral field mapped to SNOMED concept code; not extracted from referral letters or any PDF documents ▪ Risk factors: information in the Risk Factor table categorized by text pattern matching to one of the following headings: Smoking, Alcohol, Exercise, Diet, Obesity, or Psychosocial Stress. Smoking and Alcohol information further parsed (when available) into Status, Dose, Frequency, Duration, End date ▪ Family history: relationship type is identified by pattern matching as one of: Mother, Father, Daughter, Son, (Half) Sister, (Half) Brother, (Great) Grandmother, (Great) Grandfather, (Great) Aunt, (Great) Uncle, Niece, Nephew, Cousin, or None. Relationship side (Maternal or Paternal) is also identified by text pattern matching
<p>Data management organization's data transformation routines, including constructed variables</p>	<p>CPCSSN Data Dictionary⁷⁴ highlights CPCSSN-constructed variables in blue, including:</p> <ul style="list-style-type: none"> ▪ Calculated age: current year minus year of birth (<i>only as part of the national data provided to approved data users</i>) ▪ BMI: uses original BMI if available; otherwise, CPCSSN algorithms generate BMI from original height and weight data where none previously exists ▪ Deprivation category: maps dissemination areas (geographical units linked to postal code) to material and social quintiles⁷⁵ (<i>for future use in late 2019</i>) ▪ Disease case and disease case indicator tables: CPCSSN has developed and validated case definitions based on combinations of diagnoses, billing codes, medications and lab results; patients are indexed according to these definitions. Definitions for conditions currently include: diabetes, hypertension, depression, osteoarthritis, chronic obstructive pulmonary disease, dementia, parkinsonism, epilepsy¹⁸, plus chronic kidney disease, herpes zoster, and pediatric asthma <p>CPCSSN redacts several types of text that could potentially identify patients, providers, or clinics.</p> <ul style="list-style-type: none"> ▪ Non-medical words that may identify specific CPCSSN patients and providers are compared against exclusion lists, containing common words and/or allergy, diagnosis, and/or medication-specific terms to determine which are to be

	<p>redacted. The exclusion list(s) to apply depend on the type of data being deidentified. Only those names not in the exclusion list(s) will be redacted in free text.</p> <ul style="list-style-type: none"> Other identifiers, such as social insurance numbers, province health numbers, driver's license numbers, credit card numbers, Worker's Compensation Board numbers, email addresses, and phone numbers (international, North American, and local) are suppressed
Data processing validation routines	<p>During extractions, the CPCSSN data are verified by:</p> <ul style="list-style-type: none"> Comparing the number of files and data volume between extraction cycles occurring every six months Patient and provider counts Table row counts Counts for CPCSSN-specific labs and exams <p>During the data transformation, a time-stamped log-file provides an itemized description of all processing steps completed, including:</p> <ul style="list-style-type: none"> Initial and final row counts for all tables Type(s) of transformation(s) and number of data elements transformed Warnings and error messages (e.g. tables missing, processes that cannot be completed) <p>During the coding and cleaning processes, the following validation tests are performed:</p> <ul style="list-style-type: none"> Time-stamped log-file which includes an itemized list of completed processes and warnings where data should be coded or is outside allowable ranges Coder process verification tests used to compare the effect of software changes on specific coding or cleaning processes (e.g. ATC coding, exam data cleaning, deidentification) pre- and post-change End-to-end test for case definition integration that compares changes in the validity metrics (i.e. sensitivity, specificity) for CPCSSN case definitions after any changes to any part of the coding and cleaning software
Audit trail	<p>Separate time- and date-stamped log files are retained for each of the extraction, transformation, and cleaning/coding stages.</p> <ul style="list-style-type: none"> Extraction: one file per clinic; contains lists of tables and columns extracted and any warnings or errors that occurred (e.g. connection lost, table missing). An introspection table is created detailing the properties of the source database (e.g. table/column names, data types). Transformation: one file per clinic; contains software version information, file locations, initial and final row counts for all tables, type(s) of transformation(s) and number of data elements transformed, warnings and error messages (e.g. tables missing, processes that cannot be completed) Coding/Cleaning: one per PBRN; includes an itemized list of processes completed, warnings where data cannot be coded or is outside of allowed ranges

Abbreviations: ATC=Anatomical Therapeutic Chemical; BMI=Body Mass Index; CPCSSN=Canadian Primary Care Sentinel Surveillance Network; DBP=Diastolic Blood Pressure; DIN=Drug Information Number; EMR=Electronic Medical Record; ERD=Entity Relationship Diagram; ICD=International Classification of Disease; ICPC=International Classification of Primary Care; LOINC=Logical Observation Identifiers Names and Codes; NAPCRen=Northern

Alberta Primary Care Research Network; PBRN=Practice-Based Research Network; PEFr=Peak Expiratory Flow Rate; SAPCRen=Southern Alberta Primary Care Research Network; SBP=Systemic Blood Pressure; SFTP=Secure File Transfer Protocol; SNOMED=Systematized Nomenclature of Medicine; SQL=Structured Query Language.

Figure 2.1 Potential sources of bias and data quality issues in Canadian primary care EMR data (adapted from Verheij et al.⁷⁶)

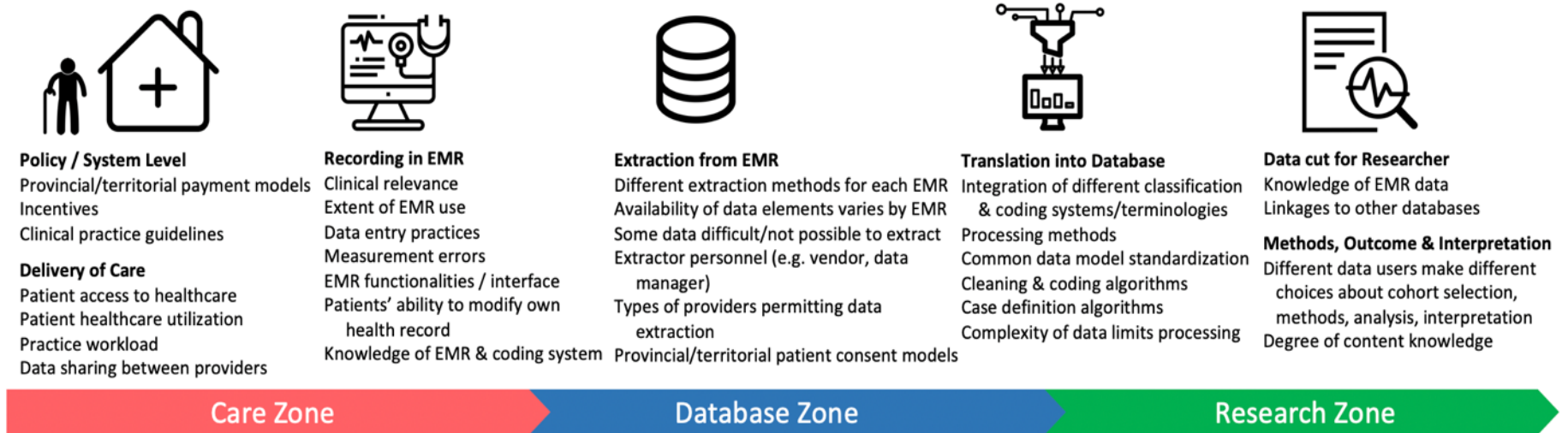
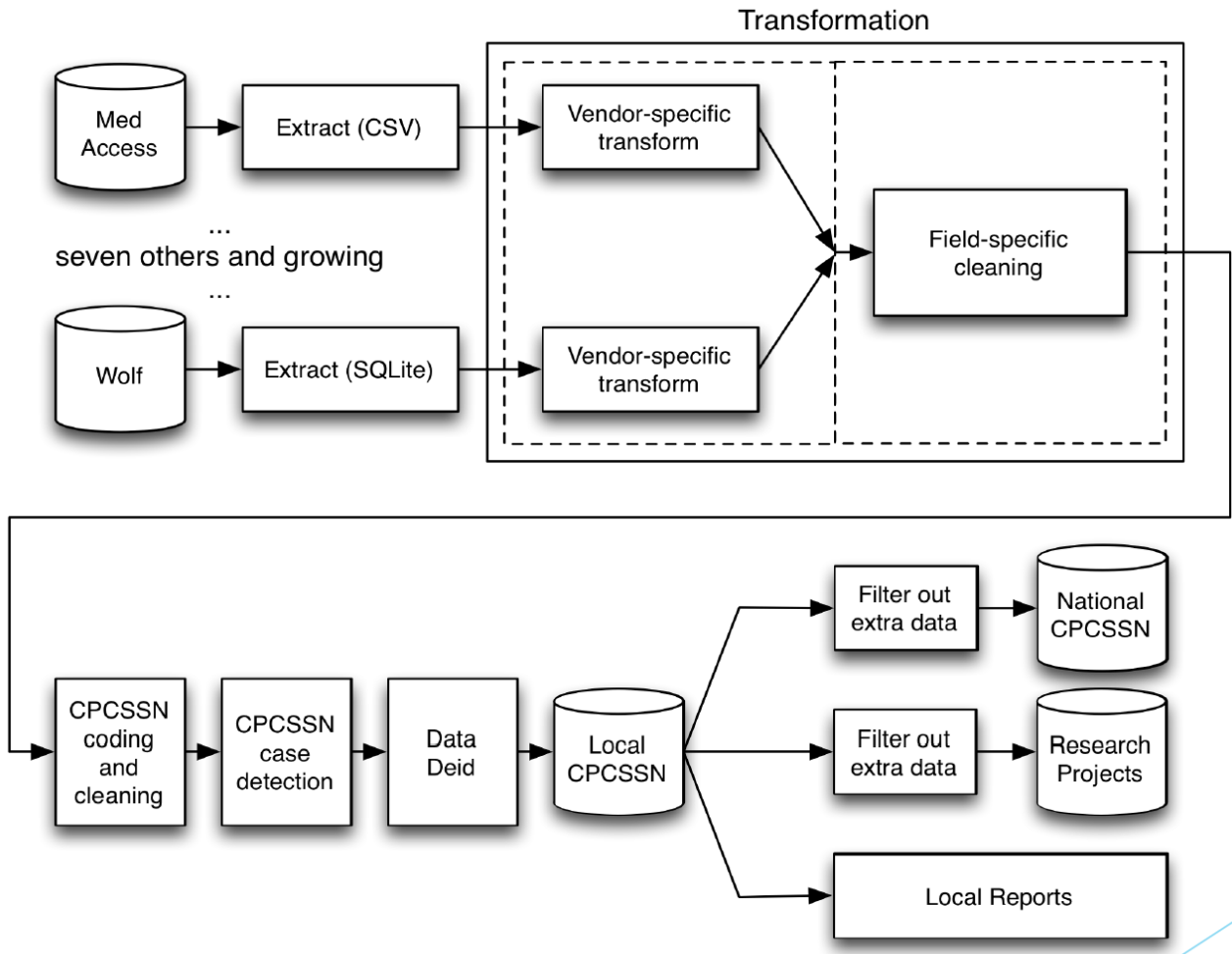


Figure 2.2 CPCSSN data pipeline



Abbreviations: CPCSSN=Canadian Primary Care Sentinel Surveillance Network; CSV=comma separated values

**CHAPTER THREE: A DATA QUALITY ASSESSMENT TO INFORM
HYPERTENSION SURVEILLANCE USING PRIMARY CARE
ELECTRONIC MEDICAL RECORD DATA FROM ALBERTA, CANADA**

This chapter is currently being revised for *BMC Public Health* (as of June 29, 2020):

Garies S, McBrien K, Quan H, Manca D, Drummond N, Williamson T. A data quality assessment to inform hypertension surveillance using primary care electronic medical record data from Alberta, Canada. *BMC Public Health*.

3.1 Abstract

Background: Hypertension is a common chronic condition affecting nearly a quarter of Canadians. Hypertension surveillance in Canada typically relies on administrative data (e.g. hospital discharge summaries; physician billing claims) and/or national surveys. Routinely-captured data from primary care electronic medical records (EMRs) have emerged as a complementary source for chronic disease surveillance and can provide longitudinal patient-level details such as sociodemographics, blood pressure, weight, prescribed medications, and behavioural risk factor information. As EMR data are generated from patient care and administrative tasks, assessing data quality is essential before using for secondary purposes. The objective of this study was to evaluate the quality of primary care EMR data from one province in Canada within the context of use for hypertension surveillance.

Methods: We conducted a cross-sectional, descriptive study using primary care EMR data collected by two practice-based research networks in the province of Alberta, Canada. There were 48,377 adults identified with hypertension from 53 primary care clinics as of June 2018. Summary statistics and visualizations were used to examine the quality of data elements considered relevant for hypertension surveillance.

Results: Patient year of birth and sex were complete, but other sociodemographic information (ethnicity, occupation, education) was largely incomplete and highly variable. Height, weight, body mass index and blood pressure were complete for the majority of patients (over 90%), but a small proportion of outlying values indicate data inaccuracies were present. Details of prescribed antihypertensive medication, such as start date, strength, dose, frequency, were mostly complete. Nearly 80% of patients had a smoking status recorded, though only 66% had useful information

(i.e. categorized as current, past, or never), and less than half had their alcohol use described; information related to amount, frequency or duration was not available.

Conclusions: Blood pressure and prescribed medications in primary care EMR data demonstrated good completeness and plausibility, and contribute valuable information for hypertension epidemiology and surveillance. The use of other clinical and sociodemographic variables may be limited due to variable completeness and suspected data errors. Developing strategies to improve these data elements at the point of data entry and after data extraction (e.g. statistical methods) is required.

3.2 Background

Hypertension is a common chronic condition, affecting more than one in five Canadians, and is associated with an increased risk of cardiovascular disease and mortality, as well as considerable economic and societal costs.⁵ Monitoring the incidence and prevalence of hypertension over time is an important part of surveillance systems and public health activities. In Canada, administrative databases, which include in-patient hospital discharges and physician billing claims, are often used to report on hypertension prevalence estimates, such as the Canadian Chronic Disease Surveillance System (CCDSS).⁷⁸ While administrative sources provide population-level data for those who have encountered the healthcare system, there are a lack of clinical details that are essential for better understanding the patient context and disease severity, including blood pressure (BP), body mass index (BMI), and lifestyle risk factors. Physical measures surveys are another commonly used source, as they obtain directly measured BP coupled with health-related interviews, as achieved by the Canadian Health Measures Survey

(CHMS).⁷⁹ However, these surveys are costly to maintain, response rates are often low, and the cross-sectional design does not allow for longitudinal follow-up.

A contemporary approach to hypertension surveillance is utilizing the clinically-generated, detailed data from electronic medical records (EMR), particularly from primary care settings where chronic conditions are largely diagnosed and managed.^{6,9} EMR adoption among Canadian family physicians is growing, with an estimated 83% now using EMRs in practice to some degree in 2018.⁸ Additionally, linkages between primary care EMR and administrative data can further enhance surveillance opportunities by providing a more complete perspective of disease manifestation and current management practices. Because EMR data are recorded to support individual patient care and administrative tasks, they may not be produced with the same standardization and rigor as research data; as such, some concern exists about their re-use for secondary purposes.³⁴ Therefore, investigations into data quality are necessary to determine whether the data are ‘fit for purpose’. Previous studies evaluating the quality of primary care EMR data in Canada have typically reported on a limited aspect of quality (e.g. completeness) or data elements^{33,68,80} or have assessed quality more broadly without focusing on a specific context for use.^{81,82} The objective of this study was to comprehensively assess the quality of primary care EMR data in Alberta, Canada within the context of hypertension.

3.3 Methods

3.3.1 Data Source

The Canadian Primary Care Sentinel Surveillance Network (CPCSSN) is a collaboration of eleven practice-based research networks (PBRN) across Canada who manage the extraction, cleaning and processing of de-identified EMR data from primary care settings.¹⁷ At present, over

1,200 primary care providers and 1.8 million patients contribute data from eight provinces and territories.⁷² National CPCSSN data have been previously used to report on the epidemiology of many conditions in primary care, such as hypertension,⁶ diabetes,²⁰ depression,²³ osteoarthritis,²² dementia,²⁴ chronic obstructive pulmonary disease,²¹ and others. The CPCSSN organization and data extraction and processing have been described elsewhere.^{17,83}

This data quality assessment utilized primary care EMR data obtained by the two PBRNs in the province of Alberta – the Northern and Southern Alberta Primary Care Research Networks (NAPCRen and SAPCRen, respectively). Because healthcare in Canada is organized and delivered separately within each province or territory, only one province (Alberta) was chosen for the data quality assessment in order to minimize variation in the data due to interprovincial differences such as healthcare delivery and practice, drug coverage, health information legislation, EMR uptake and extent of use, types of EMR systems available, and many other factors.^{7,84}

In Alberta, there were 323 providers (mostly family physicians with a small proportion of nurse practitioners and community pediatricians) participating from 53 primary care practices. This represents slightly over 5% of the total number of family physicians in Alberta.⁷³ As of June 2018, de-identified EMR data were extracted from 397,518 patients in total; this reflected approximately 9.2% of Alberta's general population of 4.3 million people.⁸⁵ The CPCSSN data has previously been found to overrepresent older adults and women, but this is typical of primary care populations.³¹

Currently, CPCSSN in Alberta extracts from five distinct EMR systems – Wolf, Med Access, Practice Solutions Suite, Accuro and Healthquest. The earliest (or 'start') date of

information in the CPCSSN database varies by clinic and by patient, depending on when a clinic first implemented their EMR system, as well as when the patient first attended the clinic.

3.3.2 Patient Sample

Adult patients (18 years and older) who had at least one primary care encounter in the previous two years (July 1, 2016 to June 30, 2018) were included, in order to establish an ‘active’ patient population. Any patient who was recorded as ‘deceased’ or ‘inactive’ in the EMR was excluded, as were any patients or providers who had explicitly requested to opt out of the CPCSSN database. The data quality assessment focused specifically on patients with hypertension who were identified using a CPCSSN-developed definition.¹⁸ The hypertension definition consisted of a combination of International Classification of Disease version nine (ICD-9) codes (401, 402, 403, 404, 405) and medications located throughout the EMR: a minimum of two physician billing codes within two years *or* any occurrence of a diagnosis in the Problem List/Profile *or* prescription for an anti-hypertensive medication (with medication criteria alone being insufficient if other specific diagnoses exist, such as heart failure or diabetes).¹⁸ The definition was validated using chart reviews as the reference standard and demonstrated good sensitivity (84.9%) and specificity (93.5%).¹⁸

3.3.3 Data Quality Assessment

The data quality assessment was a cross-sectional, descriptive evaluation guided by reporting recommendations for distributed data networks.³ Data elements were selected based on their potential use and relevancy for hypertension surveillance, as well as availability in the CPCSSN data. These included: patient demographics; physical examinations (weight, height,

body mass index [BMI], and systolic and diastolic blood pressure); anti-hypertensive medications (defined using categories of the relevant groups of Anatomical Therapeutic Chemical [ATC] codes: C02*, C03*, C07*, C08*, C09*); and risk factor records for smoking and alcohol use. Only the CPCSSN-processed/coded values were used, as these are typically the data elements that are accessible from CPCSSN for secondary purposes. A full description of all data elements can be found in the CPCSSN Data Dictionary online.⁷²

Summary statistics were reported for continuous variables, which included range, mean, and median. Proportions (restricted to the three most frequent values) and number of unique values were described for categorical variables. Missingness was reported as a proportion of patients without a recorded data element (e.g. height) or record (e.g. medication, smoking); missingness of specific items within a record was also reported (e.g. dose in medication record). Data completeness was also represented visually by clinic and EMR type.

Several temporal aspects of the data were examined – the proportion of patients who had at least one physical exam measurement (height, weight, BMI, blood pressure) recorded in the previous year (July 1, 2017 to June 30, 2018) was reported, in addition to the proportion of risk factor (i.e. smoking and alcohol) and medication records that contained a stop/end date prior to the start date. An exploration of patient-level weight values over time were visualized by plotting the difference between subsequent weight measurements and the length of time (days) between subsequent measurements for individuals with at least two weight measurements.

External validity was evaluated by comparing the most recent crude hypertension prevalence estimates from three national population-level sources: administrative data from the Canadian Chronic Disease Surveillance System (CCDSS), consisting of physician billing claims, hospitalizations and prescription drug records;⁸⁶ the Canadian Health Measures Survey (CHMS),

which defines hypertension based on standardized, direct BP measurements and health-related interviews;⁸⁷ and self-reported high BP from the Canadian Community Health Survey (CCHS).⁸⁸ Hypertension prevalence estimates from the national CPCSSN data⁶ were also used as a comparison to the regional-level (Alberta) data.

RStudio version 1.1.456 was used for the analysis, which was conducted in 2019. This study was approved by the University of Calgary's Conjoint Health Research Ethics Board and the University of Alberta's Health Research Ethics Board.

3.4 Results

In the CPCSSN data for Alberta, there were 205,364 adult patients who had at least one primary care encounter in the previous two years; of these, 48,377 patients were identified with hypertension and who were not labelled 'inactive' at the practice or deceased. Patients in the hypertension sample had a median of 8.0 years (IQR 7) of information in their record. Figure 3.1 provides a visual summary of the completeness of data for patient demographics, physical measurements, and smoking status by each of the 53 clinics and 5 EMR systems included in the data quality assessment. The data element characterization in Tables 3.1 to 3.4 provides a more in-depth examination of the quality of hypertension-related variables.

3.4.1 Patient Demographics

Birth year was complete for all patients, as was sex (with the exception of two patients). However, nearly all socio-demographic information on patients was mostly incomplete (Table 3.1). For those who had some information recorded in ethnicity, occupation, or education fields,

the data were highly inconsistent – for instance, over 3500 unique entries were recorded for occupation and more than 75 distinct entries were found for ethnicity.

3.4.2 Height, Weight, BMI

Approximately 10% of patients were missing a height or BMI value and even fewer patients were missing weight (Table 3.2). Males had a median of four measurements for height, weight, or BMI and females had five, with these measurements showing a skewed distribution. From the lower and upper ranges of the height, weight and BMI values, it appears that data errors are present. For example, female weight values ranged from 1.8 to 477 kilograms (kg), which is biologically unlikely. When plotting patient-level height and weight values (Figure 3.2), those located outside the main cluster of points visually identify specific data errors. For instance, the vertical line of points approaching 0 on the x (weight) axis might indicate a data entry error (e.g. weight entered as 10 instead of 100) or swapped height and weight values (e.g. height in metres entered in the weight field). Another observable area of atypical points was between 150 and 200 on the x (weight) axis, which potentially represents height and weight values that were entered in the wrong fields (e.g. weight=175 and height=100 recorded instead of weight=100kg and height=175cm).

Figure 3.3 investigates possible errors in weight values using successive patient-level weight measurements for those who had at least two weight values recorded in their EMR (n=39,202). It would be expected that changes in individual weight might demonstrate more variability over time (e.g. patient weight recorded 10 days apart should have minimal difference, whereas weight measurements taken several years apart might show a more significant change). Two peaks centred around 100 and -100 on the y-axis emerged as potentially problematic data:

in a relatively short time period between measurements, the difference between successive weight measurements was approximately 100 kilograms for patients clustered around those two peaks. This likely represents inconsistencies in the unit of measurement (e.g. kilograms versus pounds) for subsequent weight measurements for a given individual. However, the extent of the problem was not substantial – the majority of weight values (94.8%) occurred within two standard deviations of the central peak (mean -0.29) and at least one potential data error (i.e. outside two standard deviations) at any time was detected in the records of 18.4% of patients.

3.4.3 Blood Pressure

BP measurements were well-recorded in terms of completeness (99%) and the majority of patients (85%) had at least one measurement recorded in the previous year (Table 3.2). However, BP values at the minimum and maximum end of the range may indicate data errors (Table 3.2). These values could be biologically possible, but would be very unlikely in an outpatient setting; for instance, a systolic BP of 52 might indicate shock and a systolic BP of 290 would be an emergency event. In addition, CPCSSN also sets limits to BP values when processing the raw EMR data (50-300 mmHg for sBP; 20-200 mmHg for dBP), which would underestimate the true range of values.

Male patients had a median of 16 total BP measurements recorded in their EMR and females had slightly more (median=18). A small proportion of patients had large sums of annual BP measurements – for instance, 4.2% of females and 4.0% of males were above the 95th percentile for number of BP measurements (greater than 10) in 2017 (data not shown).

3.4.4 Hypertensive Medications

The vast majority of males (92%) and females (93%) with hypertension had at least one recorded anti-hypertensive prescription, with a median of six anti-hypertensive medication prescriptions per person (Table 3.3). The medication records themselves were fairly complete; all records contained a start date and most contained a stop date, strength, dose, frequency, duration, and count. Drug Identification Number (DIN) and ‘reason for medication’ mostly incomplete, with DIN missing in over half of medication records and ‘reason’ missing in over three-quarters of records.

3.4.5 Smoking and Alcohol Status

Within the Risk Factor section in the EMR, nearly 80% of patients had a smoking status recorded, with ‘Unknown’ and ‘Never’ as the most frequently recorded categories (Table 3.4). However, after excluding the indiscriminate ‘Unknown’ smoking status, a total of 31,976 patients (66.1%) and 68,110 records remained across three categories: ‘Current’, ‘Past’ or ‘Never’ (data not shown). Males and females had a similar number of smoking records per person (median=1; mean=3). All start and end dates were missing from the records.

More males than females had their alcohol use recorded (47% and 40%, respectively) and these records were primarily for ‘Current’ users (Table 3.4), indicating that alcohol use is likely recorded differentially between users and non-users. Patients had a mean of 2 records in their EMR (median of 1) and no records contained start or end dates.

Of note, a ‘Date Created’ field exists for both smoking and alcohol records. This field indicates when the record was created in the EMR system but does not necessarily correspond to

the start of the risk behaviour. ‘Date Created’ was present in 80.6% of smoking records and 76.6% of alcohol use records.

3.4.6 External Validity

The overall crude estimate for Alberta-specific hypertension prevalence in the CPCSSN data (23.6%) were similar to the 2014-15 physical measure survey (CHMS) (23.3%) and was also comparable to the national CPCSSN estimate (22.8%) (Table 3.5). The largest discrepancy was seen in the self-reported CCHS, with hypertension prevalence estimated at 17.7%. Male patients in the CPCSSN database had a higher hypertension prevalence (26.1%) than all other sources, while the prevalence for female patients (21.6) in the CPCSSN data was similar to the health measures survey (22.0%) and slightly lower than the CCDSS (25.6%).

3.5 Discussion

This paper describes the quality of primary care EMR data in Alberta within the context of utilization for hypertension surveillance and epidemiology. Overall, there was observable variability due to the type of EMR system, between clinics, and among the data elements themselves. As this assessment focused on patients with hypertension, it was not surprising to see blood pressures and prescribed medication records that were largely complete and contained minimal outliers; these data constitute a particularly valuable contribution for surveillance purposes, given that BP and prescribing information are not available in administrative data or are limited (i.e. cross-sectional) in survey data. Although these data cannot confirm whether a patient has filled their prescription or is adherent, the information within the medication records

are relatively complete and can be used to approximate persistence/adherence, for example, by calculating medication possession ratio or using similar methods.⁸⁹

Other information, such as sociodemographic, height, weight, and risk factor information, were more inconsistent and less complete. Although achieving 100% completeness for all data elements may not be realistic, it is not unreasonable to aim for near complete information for these data elements at the point of care. Cardiovascular disease guidelines suggest that smoking status should be updated on a regular basis and given that screening is often risk based, information about alcohol use, height, weight, BMI, and ethnicity are particularly important to document for a hypertensive cohort.⁹⁰ However, distinguishing between data that are missing due to inadequate data entry or as a result of not extracting the data is difficult. One significant challenge when addressing poor data quality is determining the source of the issue – for instance, missing data may be due to the unavailability of these fields in certain EMR systems (in which case, missingness will always exist); patients might not be asked about specific topics, such as alcohol use or ethnicity, or they may decline to answer; lastly, the CPCSSN processes may omit extraction from certain fields of the EMR either deliberately (e.g. identifiable fields or physician notes) or unintentionally (e.g. if an EMR system upgrade changes the names of data elements, which would subsequently affect the CPCSSN extraction code). Identifying true inaccuracies in the data are similarly problematic; this may be possible for some data elements through a chart review, with particular attention to the detailed physician notes and scanned documents (i.e. specialist letters, diagnostic imaging) that are not currently captured in the CPCSSN data. However, this is a time-intensive method and the structured EMR fields are likely to contain the same errors and omissions as the CPCSSN data. Beyond this, confirming with or measuring

patients directly to verify data elements in the EMR would most accurately reveal true data errors but this method is also the least feasible.

Therefore, the most appropriate strategies for preventing and mitigating EMR data quality issues should be multifaceted and involve a variety of settings. CPCSSN has largely taken a post-extraction analytic approach to data improvement. This includes extensive cleaning and coding algorithms, the development and validation of case definitions for various conditions that are made available as part of the database,^{18,65–67} and exploration of more advanced techniques like natural language processing⁹¹ and machine learning.⁶⁶ As an example, CPCSSN is currently developing a pattern-matching algorithm that aims to enhance the completeness and accuracy of smoking records. In the raw or original EMR data, some additional information related to smoking, such as frequency of tobacco use and quantity of tobacco units consumed (e.g. cigarettes / cigars, packs), is present but primarily in unstructured, lengthy text strings that is not useful for analysis, may also contain identifiable patient information, and is therefore not currently available to researchers. The pattern-matching algorithm is designed to extract only smoking-related information from the free text and categorize the record into a defined smoking status, leading to more available coded data for researchers to access.

A number of other strategies have been shown to improve the completeness and accuracy of EMR data – some occur at the practice level, such as employing a dedicated data entry clerk⁴³ or providing data quality audit and feedback reports to clinicians.⁴² Other initiatives require more substantial resources and uptake, such as national EMR content standards,⁵² developing EMR interfaces that are easier to navigate and contain more structured fields, and promoting financial or other incentives for ‘meaningful EMR use’.³⁹

In the future, routine linkage to other data sources, like administrative health data, could enhance quality by providing a mechanism to verify certain aspects of EMR data and expand the breadth of information about individual patients throughout the broader healthcare system.

3.5.1 Limitations

This paper provides a quality assessment of select CPCSSN data elements deemed to be important for hypertension surveillance or research, but it was not possible to examine and report on all variables contained in the CPCSSN database in a single manuscript. It was also not feasible to quantify the true accuracy of data elements, other than appraising the plausibility of values through descriptive means. Secondly, during the CPCSSN processing and data transformation stages, restrictions are introduced for out-of-bounds values and thus, the summary statistics presented in this paper may not reflect the full variation of values originating from the source EMR. In addition, any changes or improvements made to the CPCSSN processing may result in slight differences in the CPCSSN EMR database between each extraction cycle. Thirdly, although the CPCSSN definition for hypertension demonstrated high sensitivity and specificity, a potential for misclassification still exists. This may have underestimated the number of patients with hypertension or produced a patient sample that is biased towards a greater severity of illness. Lastly, the quality was described specifically for CPCSSN data from Alberta and within the context of hypertension. This is not a population-level data source and only constitutes a sample of participating providers and patients who have sought care. Thus, the overall findings may not be representative of the wider Alberta population or for other provinces or territories that participate in CPCSSN, and may also differ in other disease-based contexts. However, CPCSSN has developed uniform data extraction, processing, and standardization

methods across the country, which may allow for other regional networks to compute the same data quality assessment for comparison.

3.6 Conclusion

Primary care EMR data are a valuable data source for hypertension surveillance or within an epidemiological context. The high-quality and longitudinal blood pressure and prescribed antihypertension medication data are particularly useful, as these types of data are not found in traditional administrative databases. Other data elements, such as sociodemographics, physical examination values, and risk factor information, exhibited variation in quality. These data elements may be less useful in their current state, but offer promising value in the future once data quality issues can be addressed through additional pre- or post-extraction solutions.

Table 3.1 Missingness and summary statistics for patient demographic information*

Data Element	Summary	Males N=23,486 (48.5%)	Females N=24,889 (51.5%)
Year of birth	No. patients missing, n (%)	0 (0)	0 (0)
	Range	1915-1998	1914-1998
	Mean (SD)	1954 (13.4)	1952 (14.4)
	Median (IQR)	1954 (17)	1951 (19)
Ethnicity	No. patients missing, n (%)	22,511 (95.8)	23,788 (95.6)
	Unique values, n	76	83
	3 most frequent values, n (%)	“Caucasian”: 623 (63.9) “Aboriginal”: 103 (10.6) “Canadian”: 64 (6.6)	“Caucasian”: 588 (53.4) “Aboriginal”: 145 (13.2) “Canadian”: 99 (9.0)
Occupation	No. patients missing, n (%)	16,810 (71.6)	18,423 (74.0)
	Unique Values, n	3628	3536
	3 most frequent values, n (%)	“Retired”: 1080 (16.2) “Truck driver”: 125 (1.9) “Farmer”: 103 (1.5)	“Retired”: 1250 (19.3) “Homemaker”: 81 (1.3) “Teacher”: 80 (1.2)
Education	No. patients missing, n (%)	22,866 (97.4)	24,262 (97.5)
	Unique Values, n	10	12
	3 most frequent values, n (%)	“University”: 184 (22.7) “High School”: 176 (28.4) “College”: 89 (14.4)	“University”: 188 (30.0) “High School”: 175 (27.9) “College”: 151 (24.1)

Abbreviations: IQR=interquartile range; SD=standard deviation

*Sex was missing for 2 patients

Table 3.2 Missingness and summary statistics for physical measurement values

Data Element	Summary	Males N=23,486	Females N=24,889
Height (cm)	No. patients missing, n (%)	2391 (10.2)	2345 (9.4)
	Patients with a measurement in previous year, n (%)	11,604 (49.4)	12,562 (50.5)
	Range	41.9-229.6	37.2-229.0
	Mean (SD)	174.6 (8.2)	160.8 (7.5)
	Median (IQR)	175.0 (10.0)	161.0 (9.2)
	Median number of total measurements per patient, n	4.0	5.0
Weight (kg)	No. patients missing, n (%)	1791 (7.6)	1714 (6.9)
	Patients with a measurement in previous year, n (%)	12,165 (51.8)	13,241 (53.2)
	Range	1.0-500.0	1.8-477.0
	Mean (SD)	104.9 (41.5)	90.4 (39.0)
	Median (IQR)	94.0 (30.3)	80.0 (33.0)
	Median number of total measurements per patient, n	4.0	5.0
BMI	No. patients missing, n (%)	2422 (10.3)	2316 (9.3)
	Patients with a measurement in previous year, n (%)	11,979 (51.0)	13,062 (52.5)
	Range	5.1-199.9	5.0-200.0
	Mean (SD)	31.8 (9.5)	32.2 (11.1)
	Median (IQR)	30.2 (7.4)	30.1 (10.1)
	Median number of total measurements per patient, n	4.0	5.0
Systolic blood pressure (mmHg)	No. patients missing, n (%)	226 (1.0)	264 (1.1)
	Patients with a measurement in previous year, n (%)	19,898 (84.7)	21,069 (84.6)
	Range	52.0-266.0	54.0-290.0
	Mean (SD)	134.4 (17.2)	134.6 (17.7)
	Median (IQR)	133.0 (22.0)	133.0 (23.0)
	Median number of total measurements per patient, n	16.0	18.0
Diastolic blood pressure (mmHg)	No. patients missing, n (%)	226 (1.0)	264 (1.1)
	Patients with a measurement in previous year, n (%)	19,898 (84.7)	21,069 (84.6)
	Range	30.0-188.0	30.0-200.0
	Mean (SD)	80.4 (11.5)	78.8 (11.0)
	Median (IQR)	80.0 (16.0)	80.0 (15.0)
	Median number of total measurements per patient, n	16.0	18.0

Abbreviations: BMI=body mass index; cm=centimetres; IQR=interquartile range; kg=kilograms; mmHg=millimeter of mercury; SD=standard deviation

Table 3.3 Missingness and summary statistics for anti-hypertensive medications

Data Element	Summary	Males N=23,486	Females N=24,889
Anti-hypertensive medication	No. patients with prescription at any time, n (%)	21,544 (91.7)	23,184 (93.1)
	Mean prescriptions per patient, n (SD)	12.1 (18.0)	12.0 (17.4)
	Median prescriptions per patient, n (IQR)	6.0 (10)	6.0 (11)
Medication name	No. records missing med name, n (%)	0 (0)	0 (0)
	Unique values, n	86	87
	3 most frequent values, n (%)	‘Ramipril’: 33,109 (12.7) ‘Amlodipine’: 27,772 (10.7) ‘Hydrochlorothiazide’: 26,238 (10.1)	‘Hydrochlorothiazide’: 40,537 (14.6) ‘Amlodipine’: 26,892 (9.7) ‘Ramipril’: 25,721 (9.3)
Start date	No. records missing start date, n (%)	0 (0)	0 (0)
	Unique values, n	5471	5604
Stop date	No. records missing stop date, n (%)	34,404 (13.2)	39,686 (14.3)
	Unique values, n	5664	5751
	Stop date occurs before start date, n (%)	2272 (0.9)	2364 (0.9)
Drug Identification Number (DIN)	No. records missing DIN, n (%)	135,564 (52.1)	150,923 (54.3)
	Unique values, n	1175	1238
	3 most frequent values, n (%)	‘326844’: 5264 (4.2) ‘878928’: 5262 (4.2) ‘2123282’: 4324 (3.5)	‘326844’: 7091 (5.6) ‘878928’: 6265 (4.9) ‘2123282’: 3452 (2.7)
Strength	No. records missing strength, n (%)	22,914 (8.8)	32,396 (11.7)
	Unique Values, n	95	99
	Median (IQR)	12 (35)	20 (35)
Dose	No. records missing dose, n (%)	5428 (2.1)	7387 (2.7)
	Unique values, n	2048	2730
	3 most frequent values, n (%)	‘1’: 202,129 (79.3) ‘0.5’: 15,563 (6.1) ‘2’: 8635 (3.4)	‘1’: 208,017 (77.0) ‘0.5’: 18,644 (6.9) ‘2’: 7829 (2.9)
Frequency	No. records missing frequency, n (%)	24,013 (9.2)	33,261 (12.0)

	Unique values, n	133	117
	3 most frequent values, n (%)	‘QD’: 140,843 (59.6) ‘OD’: 45,494 (19.3) ‘BID’: 32,140 (13.6)	‘QD’: 145,261 (59.5) ‘OD’: 52,094 (21.3) ‘BID’: 27,764 (11.4)
Duration count	No. records missing duration, n (%)	33,013 (12.7)	38,379 (13.8)
	Unique values, n	3507	3692
	3 most frequent values, n (%)	‘3’: 51,704 (22.8) ‘100’: 50,550 (22.2) ‘90’: 21,444 (9.4)	‘3’: 51,928 (21.7) ‘100’: 49,245 (20.6) ‘90’: 20,498 (8.6)
Duration unit	No. records missing duration unit, n (%)	39,628 (15.2)	47,796 (17.2)
	Unique values, n	5	5
	3 most frequent values, n (%)	‘Day’: 139,344 (63.2) ‘Month’: 72,838 (33.0) ‘Week’: 5893 (2.7)	‘Day’: 145,621 (63.4) ‘Month’: 77,349 (33.7) ‘Week’: 4025 (1.8)
Dispensed count	No. records missing dispensed count, n (%)	28,376 (10.9)	33,779 (12.2)
	Unique values, n	3688	3732
	Median (IQR)	100 (90)	91 (95)
Dispensed form	No. records missing dispensed form, n (%)	74,277 (28.5)	93,218 (33.6)
	Unique values, n	3	5
	3 most frequent values, n (%)	‘Tab’: 158,009 (84.9) ‘Capsule’: 27,994 (15.1) ‘Bottle’: 1 (0)	‘Tab’: 161,436 (87.6) ‘Capsule’: 22,859 (12.4) ‘Bottle’: 3 (0)
Reason for medication	No. records missing reason, n (%)	199,924 (76.8)	210,112 (75.7)
	Unique values, n	281	322
	3 most frequent values, n (%)	‘Hypertension, CHF’: 21,278 (35.3) ‘Hypertension, Angina’: 11,743 (19.5) ‘Hypertension’: 7711 (12.8)	‘Hypertension, CHF’: 20,928 (31.0) ‘Hypertension, Angina’: 12,268 (18.2) ‘Hypertension’: 10,337 (15.3)

Abbreviations: IQR=interquartile range; SD=standard deviation

Table 3.4 Missingness and summary statistics for risk factor records

Data Element	Summary	Males N=23,486	Females N=24,889
Smoking status	No. patients missing smoking record, n (%)	4716 (20.1)	5413 (21.7)
	Unique values, n	4	4
	3 most frequent values in all records, n (%)	“Unknown”: 17,942 (33.8) “Never”: 16,760 (31.6) “Current”: 11,799 (22.2)	“Never”: 20,018 (38.6) “Unknown”: 18,812 (36.3) “Current”: 7658 (14.8)
	Mean number of smoking records per patient, n (SD)	2.8 (4.2)	2.7 (3.6)
	Median number of smoking records per patient, n (IQR)	1.0 (2.0)	1.0 (2.0)
Smoking start date	No. records missing start date, n (%)	52,999 (99.9)	51,780 (99.9)
	Smoking end date occurs before start date, n (%)	0 (0)	0 (0)
Alcohol use status	No. patients missing alcohol record, n (%)	12,372 (52.7)	15,034 (60.4)
	Unique values, n	5	7
	3 most frequent values in all records, n (%)	‘Current’: 7228 (93.6) ‘Past’: 479 (6.2) ‘Unknown’: 11 (0.1)	‘Current’: 6068 (94.8) ‘Past’: 220 (3.4) ‘Unknown’: 108 (1.7)
	Mean number of alcohol records per patient, n (SD)	2.0 (2.2)	2.2 (2.6)
	Median number of alcohol records per patient, n	1.0 (1.0)	1.0 (1.0)
Alcohol use start date	No. records missing start date, n (%)	21,939 (100)	22,127 (100)
	Alcohol use end date occurs before start date, n (%)	0 (0)	0 (0)

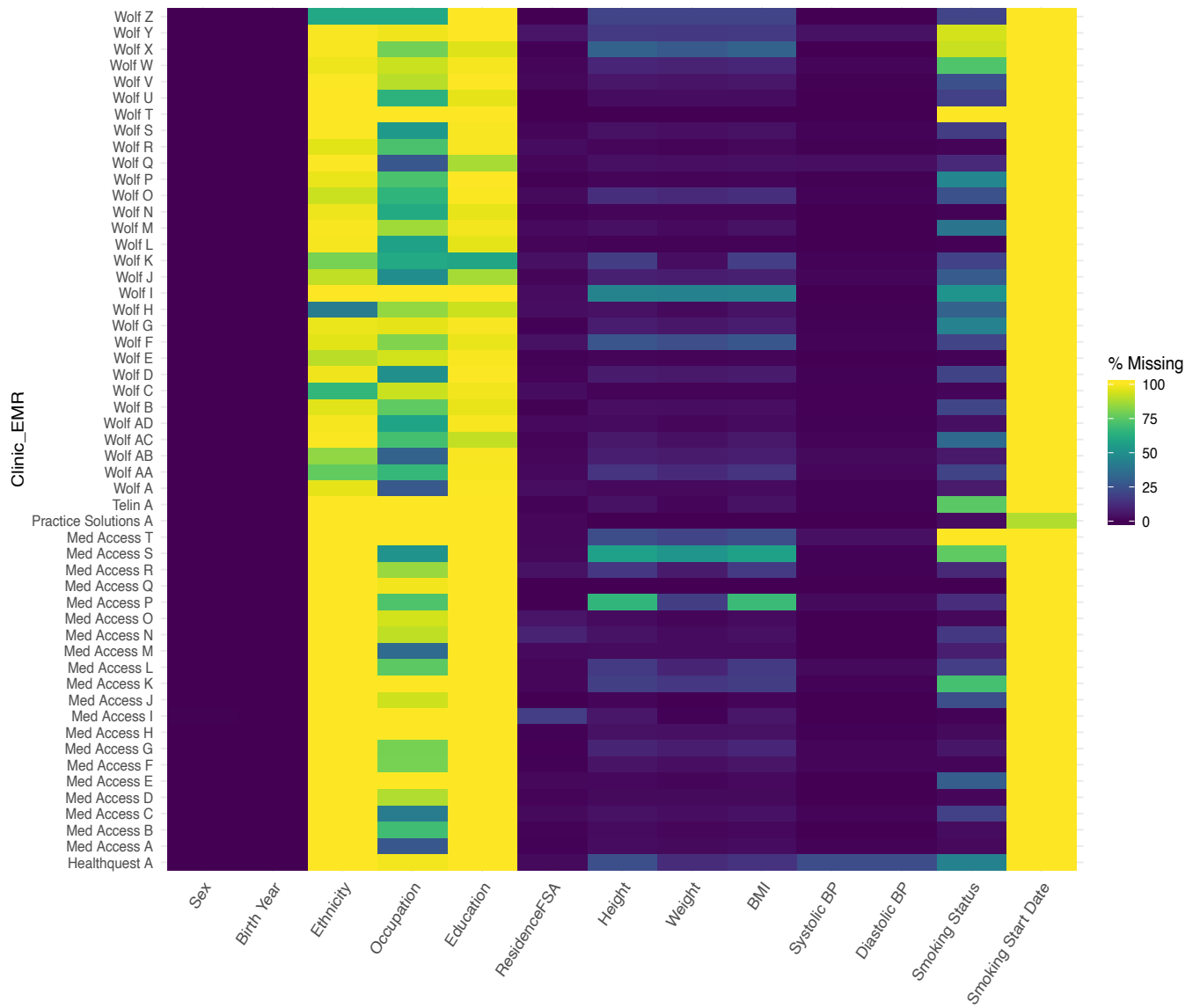
Abbreviations: IQR=interquartile range; SD=standard deviation

Table 3.5 Prevalence comparisons of adults with hypertension in various data sources

Data Source	Data Source Type	Year	Sample	Overall (crude %)	Males (crude %)	Females (crude %)
CPCSSN (Alberta-specific)	Primary care EMR data	2018	Albertans 18 years & older	23.6	26.1	21.6
National CPCSSN data ⁶	Primary care EMR data	2012	Canadians 18 years & older	22.8	24.1	21.9
CCDSS ⁸⁶	Administrative data	2015	Canadians 20 years and older	25.4	25.2	25.6
CHMS ⁸⁷	Physical measures survey	2014-15	Canadians 20-79 years	23.3	24.3	22.0
CCHS ⁸⁸	Self-reported survey	2016	Canadians 12 years and older	17.7	18.3	17.1

Abbreviations: CCDSS=Canadian Chronic Disease Surveillance System; CCHS=Canadian Community Health Survey; CHMS=Canadian Health Measures Survey; CPCSSN=Canadian Primary Care Sentinel Surveillance Network; EMR=electronic medical record

Figure 3.1 Summary of completeness of select EMR data elements by EMR type and clinic for patients with hypertension

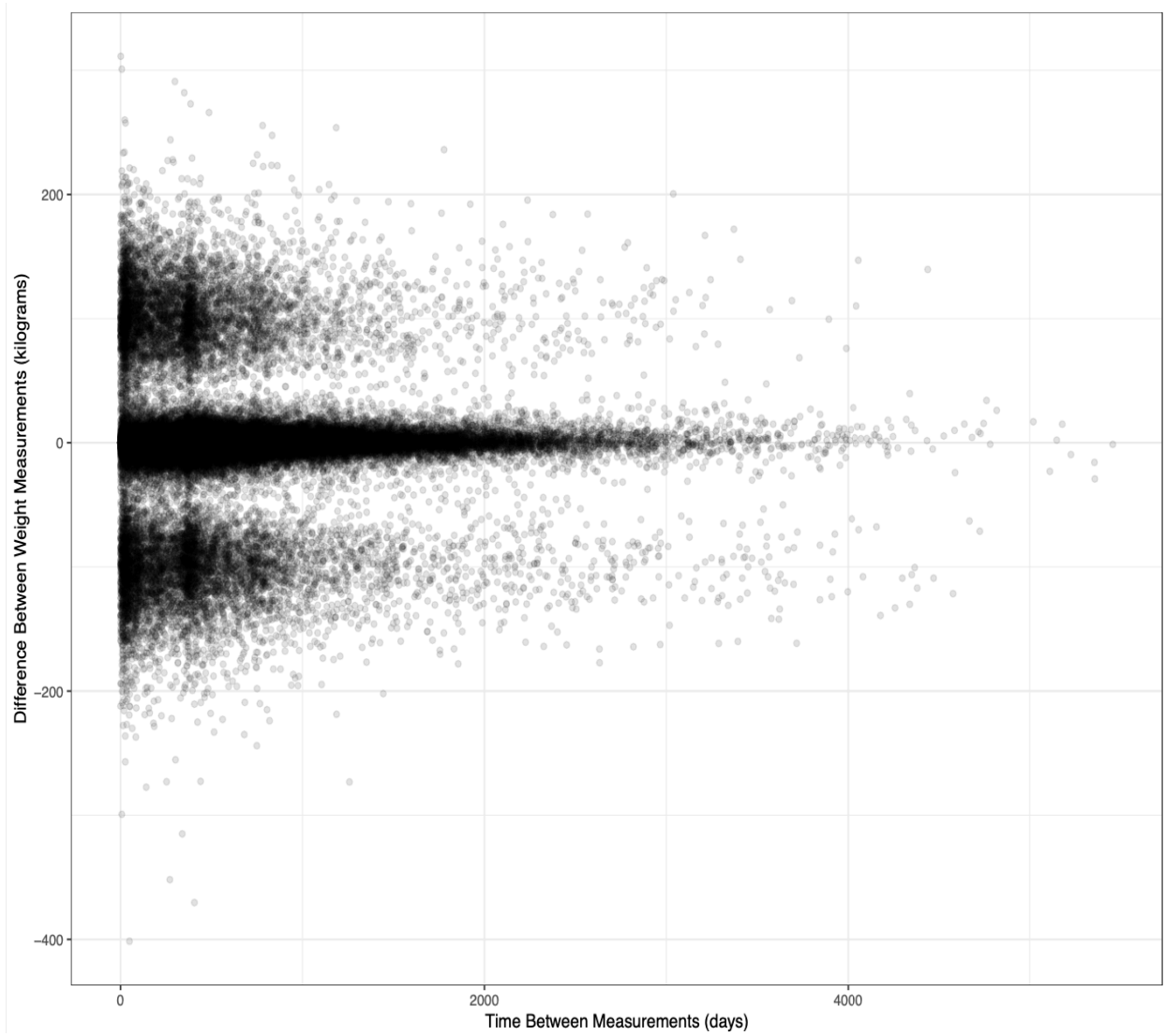


Abbreviations: BMI=body mass index; BP=blood pressure; EMR=electronic medical record; FSA=forward sortation area

Figure 3.2 Paired height and weight measurements in patients with hypertension



Figure 3.3 Differences in subsequent patient weight measurements throughout time in patients with hypertension



Note: Patients with two or more weight values recorded at any time in their EMR (n=39,202).

**CHAPTER FOUR: METHODS TO IMPROVE THE QUALITY OF
SMOKING RECORDS IN A PRIMARY CARE EMR DATABASE:
EXPLORING MULTIPLE IMPUTATION AND PATTERN-MATCHING
ALGORITHMS**

This chapter is published as:

Garies S, Cummings M, Quan H, McBrien K, Drummond N, Manca D, Williamson T.
Methods to improve the quality of smoking records in a primary care EMR database:
exploring multiple imputation and pattern-matching algorithms. *BMC Medical
Informatics and Decision Making* 2020; 20:54.

4.1 Abstract

Background: Primary care electronic medical record (EMR) data are emerging as a useful source for secondary uses, such as disease surveillance, health outcomes research, and practice improvement. These data capture clinical details about patients' health status, as well as behavioural risk factors, such as smoking. While the importance of documenting smoking status in a healthcare setting is recognized, the quality of smoking data captured in EMRs is variable. This study was designed to test methods aimed at improving the quality of patient smoking information in a primary care EMR database.

Methods: EMR data from community primary care settings extracted by two regional practice-based research networks in Alberta, Canada were used. Patients with at least one encounter in the previous two years (2016-2018) and having hypertension according to a validated definition were included (n=48,377). Multiple imputation was tested under two different assumptions for missing data (smoking status is missing at random and missing not-at-random). A third method tested a novel pattern matching algorithm developed to augment smoking information in the primary care EMR database. External validity was examined by comparing the proportions of smoking categories generated in each method with a general population survey.

Results: Among those with hypertension, 40.8% (n=19,743) had either no smoking information recorded or it was not interpretable and considered missing. Those with missing smoking data differed statistically by demographics, clinical features, and type of EMR system used in the clinic. Both multiple imputation methods produced fully complete smoking status information, with the proportion of current smokers estimated at 25.3% (data missing at random) and 12.5% (data missing not-at-random). The pattern-matching algorithm classified 18.2% of patients as current smokers, similar to the population-based survey (18.9%), but still resulted in missing

smoking information for 23.6% of patients. The algorithm was estimated to be 93.8% accurate overall, but varied by smoking status category.

Conclusion: Multiple imputation and algorithmic pattern-matching can be used to improve EMR data post-extraction but the recommended method depends on the purpose of secondary use (e.g. practice improvement or epidemiological analyses).

4.2 Background

There has been a transformative shift towards the uptake of electronic medical record (EMR) systems in primary care settings and the subsequent re-use of these data for a variety of secondary purposes. In Canada, the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) has established a source of de-identified, patient-level primary care EMR data that has been used for health research, disease surveillance, and quality improvement.¹⁷ This has also included the development of extensive processes and algorithms to extract, transform, and standardize the raw EMR data generated from nearly two million patients in seven Canadian provinces or territories.^{17,72}

One distinct advantage of the CPCSSN data is the availability of behavioural risk factor information (such as smoking, alcohol use, diet and exercise) coupled with clinically-verified health measures and outcomes. However, risk factor data are often entered in unstructured fields or as free text in different areas of the EMR, which also varies by the type of EMR system being used. This creates difficulties during the data extraction process, as well as analytic challenges for data users, and also impacts the quality of these data in terms of completeness and accuracy. For instance, more than one-third of all patients over 16 years of age in the national CPCSSN

database were missing smoking information, with biases likely introduced due to how smoking status was captured and for whom.³³

When conducting observational studies using EMR data, it is not uncommon to encounter missing data. Part of the challenge is determining whether the data are missing at random (where the probability of missingness depends on other observed information or data) or missing not-at-random (where the probability of missingness depends on unobserved or unknown information),⁹² in addition to deciding on the most appropriate technique to use for analysis when data are missing. For some situations, it may not be recommended to use a complete case analysis (i.e. including only those who do not have missing data), as this method can reduce sample size and statistical power, lower precision of results, and produce biased findings.⁹² Multiple imputation (MI) has been touted as a more accurate and less biased method, as missing data are imputed based on the distributions and variability of other data elements in the sample.⁹² The use of MI to improve the recording of smoking data has been previously demonstrated in a large U.K. general practitioner database, where the authors observed that smoking data were likely missing non-randomly and had identified potential misclassification of ex-smokers who were recorded as non-smokers, particularly if a substantial period of time had passed since quitting or if quitting occurred at a younger age.^{93,94}

The objective of this study is to explore methods aimed at improving the completeness of patient smoking status in the CPCSSN database, specifically focused on patients with hypertension. Two methods will test MI using different scenarios for data missing at random (MAR) or missing not at random (MNAR). The third method will test a newly-developed CPCSSN-based classifier for assigning individual smoking status based on additional free text in the smoking record.

4.3 Methods

4.3.1 Data Source

This study used CPCSSN data collected by two practice-based research networks in Alberta, Canada (Northern Alberta Primary Care Research Network [NAPCRen], Southern Alberta Primary Care Research Network [SAPCRen]). Over 320 community-based providers (mostly family physicians with a small proportion of nurse practitioners and pediatricians) contribute de-identified EMR data from close to 400,000 patients across Alberta. Healthcare is publicly funded in Canada and typically administered by each province or territory. There is no mandatory patient registration in primary care, although patient rostering is increasingly undertaken in Alberta.

The CPCSSN data contain nearly all information entered into the patient EMR, including demographics, patient profile (e.g., medical diagnoses, history), physical measurements (e.g., blood pressure, height, weight), risk factor information, laboratory results, prescribed medications, medical procedures, referrals, and physician billing claims.¹⁷ Additional details about the regional extraction, processing and transformation for NAPCRen and SAPCRen have been described elsewhere.⁸³

4.3.2 Sample

This analysis focused on a cohort of patients with hypertension, as guidelines recommend that tobacco use status should be updated on a regular basis for these patients.⁹⁰ Patients were identified as hypertensive if they met the validated definition developed by CPCSSN, which includes a combination of diagnostic codes related to hypertension and antihypertensive medications.¹⁸ This definition demonstrated good validity as compared to the reference standard

chart review (sensitivity=84.9%; specificity=93.5%).¹⁸ Patients with at least one encounter with their primary care provider within the previous two years (July 1, 2016 to June 30, 2018) were included; this is considered an acceptable indicator of ‘active’ patients in a practice, as the large majority of individuals visit a family physician at least once in a two-year period.⁹⁵ Those who were explicitly recorded as deceased or inactive according to their EMR status were excluded. Data from the entire history of the patient EMR were used for the analysis; the earliest timepoint varies by patient and depends on when a clinic integrated their EMR system into practice and when the patient first attended the clinic.

4.3.3 Defining Smoking Status

The smoking record for each patient found in the Risk Factor table of the CPCSSN database was used. Patient smoking status can be first documented at any time and updated as often or as infrequently as chosen by the patient and provider. For those patients who had more than one smoking status documented at any time throughout the history of their EMR, the most recent status was used according to the date the record was created in the clinic EMR system. Each risk factor record is comprised of several fields including: a unique CPCSSN identifier for each patient, date on which the record was entered into the EMR, risk factor name, risk factor status, and risk factor values (i.e. any quantitative or extraneous information). Records with the risk factor name coded as ‘Smoking’ or ‘Unknown’ were selected for this analysis. The ‘Status’ of these risk factor records contained the original EMR data and consisted of four categories: current, past, never, and unknown. For the imputation, all unknown statuses were recoded as missing data.

4.3.4 Defining Demographic and Clinical Variables

The relevant patient characteristics and clinical values included in the multiple imputation were sex, age, rural or urban residence, body mass index (BMI), systolic blood pressure, diabetes, number of clinical encounters, and type of EMR system used in the practice. Patient residence was considered rural if the second character of the postal code was 0 and urban if the second character was any digit except 0. Most recently recorded blood pressure and body mass index (BMI) were used. BMI was included if it was deemed to be within a plausible range (between 12 and 70 kg/m²) and if the difference between subsequent per-patient measurements was within 10 kg/m². Patients were indexed with diabetes and/or chronic obstructive pulmonary disease (COPD) according to validated CPCSSN definitions. The diabetes definition had a sensitivity of 95.6% and specificity of 97.1%; COPD demonstrated a sensitivity of 82.1% and specificity of 97.3%.¹⁸ The number of clinical encounters was a sum of visit dates documented in the EMR for each patient. The earliest patient encounter date in the CPCSSN database varies, depending on when a clinic first implemented an EMR system and when the patient first attended.

4.3.5 Methods for Improving Completeness

Three methods for improving the completeness of smoking data were tested. The first two utilized multiple imputation (MI) – this is an analytical technique for improving missing data and can be readily implemented using currently available statistical software (e.g. R). MI first generates many different datasets with the missing values imputed based on the distribution of missing data in the original dataset; then, analyses are conducted on each dataset separately and lastly, the results are pooled to generate a multiple imputed estimate.⁹² The advantage of MI is

that it allows uncertainty in the missing data to be accounted for and has been shown to result in more accurate standard errors and less biased outcomes compared to a complete case analysis.⁹²

MI was applied under two different missing data assumptions, similar to Marston et al.'s previous work using a UK general practice EMR database.⁹⁴ The first MI method considered smoking status to be 'missing at random' (MAR) and missing data for three smoking categories, current, past, and never, were imputed. The second MI method assumes that all current smokers have been recorded (given the importance of smoking as a significant risk factor for hypertensive patients) and thus, smoking status is 'missing-not-at-random' (MNAR). Here, only the categories of past or never smokers will be imputed for those missing smoking status data.

The third method used a classifier algorithm recently developed by a CPCSSN data manager (MC). The raw EMR data that are used to create the CPCSSN national risk factor data set, comprising approximately one million smoking records in both official languages (English and French), were used to develop text-matching patterns for calculating patient smoking status. These raw data are broken into three text fields: original Name, original Value (e.g., 10 packs/day), and original Status, where 'original' refers to the raw data from the EMR. For records that are coded as 'smoking', risk factor status is coded to one of the following categories: current, never, past, not current, and unknown. A status of 'not current' means it is unclear whether the patient is a past smoker or has never smoked (e.g., 'non-smoker'). Lastly, records that match either zero or multiple statuses are coded as 'Unknown'.

The smoking status classifier was created following manual review of the top 2,000 most frequently occurring smoking records in the national CPCSSN database, comprising approximately 75% of the data. For each record, the true smoking status was determined after review of the text in the all fields related to smoking status in the CPCSSN database (Name,

Value, and Status). This text was then used to generate a unique text-pattern that identified the true status. Once all 2,000 records were reviewed, a random sample of 1,000 records from the remaining 25% of the data set were then analysed to validate and improve pattern-matching accuracy. For each status, a simplified set of all text-patterns, plus additional logic, was then collected to create the smoking status classifier. Figure 4.1 displays a flow chart of the classifier's operation. It analyzes each record sequentially by field, with data in the Status field having precedence over Value, and Value over Name, as shown in Figure 4.1. Entries that contain relationship words (e.g., aunt, father) are ignored by the classifier, as they are usually family history or other non-risk factor data that typically cause patient smoking status to be misclassified.

4.3.6 Analysis

Patient demographic and clinical characteristics were reported as proportions for categorical variables, mean with standard deviation (SD) for normally distributed continuous variables, and median with lower and upper quartiles for non-parametric data. Comparisons of characteristics between patients with and without smoking records were conducted using a chi-squared test for categorical variables, one-way test for continuous variables, and Wilcoxon test for non-parametric data.

For the multiple imputation in Method 1 and 2, the MICE package in R was used.⁹⁶ This analytic tool uses a fully conditional specification to impute each variable by specifying an imputation model that best suits the variable type. For instance, a Bayesian polytomous regression model was used for the unordered, categorical smoking status data. Forty imputed datasets were generated with 5 iterations taken to impute the missing values.

To obtain error estimates for the MI methods, a simulation was conducted for patients with complete smoking status data. Missing smoking data were randomly introduced for 40.8% of the complete case group (the same proportion of missing smoking status data in the total sample of 48,377).

The classifier algorithm used by Method 3, programmed in Python, was applied to the hypertensive dataset and a ‘calculated’ Status was generated for each smoking record. Method 3 was then validated by manually assigning a smoking status to all unique records, based on the original Status and Value data, while blinded to the result of the algorithmic-calculated Status. Since, for the hypertensive data set, the original Name data only contain one string (‘Smoking’), they were not reviewed during algorithm assessment.

External validity was estimated using Alberta-specific data from the 2017 Canadian Tobacco, Alcohol, and Drugs Survey (CTADS).⁹⁷ CTADS is a biennial cross-sectional telephone survey in Canadians ages 15 years and older conducted by Statistics Canada.⁹⁷ Results are available by province and therefore, the smoking status of Alberta respondents was used (n=1444 survey respondents weighted to a population estimate of 3,478,000).

All data analysis was performed in RStudio version 1.1.456. This study was approved by the Conjoint Health Research Ethics Board at the University of Calgary and the Health Research Ethics Board at the University of Alberta.

4.4 Results

The flow of patients into the study cohort is summarized in Figure 4.2. There were 48,377 adult patients with hypertension identified in the CPCSSN data from Alberta within a two-year contact period who were not deceased or inactive. Of these, 28,634 (59.2%) had a

smoking status recorded and available in the database, with approximately 20% having more than one smoking status entry recorded throughout the history of their EMR. For smoking entries with an associated date of entry, nearly 80% were recorded or updated within the previous two years (as the 'most recent' status was chosen for inclusion). Table 4.1 summarizes demographic, clinical, and practice characteristics for those with and without smoking information. Those who were missing smoking status information (n=19,743) were more likely to be older (p=0.001), female (p=0.001), have higher systolic (p<0.001) and diastolic (p=0.027) blood pressure, lower BMI (0.001), less co-morbid diabetes and COPD (p<0.001), reside in an urban setting (p<0.001), and have fewer total clinical encounters (p<0.001). While the differences in sex, age, BMI, and blood pressure between the two groups were statistically significant, they were small in absolute terms and may not be clinically meaningful.

The two groups also differed significantly by the type of EMR system (Wolf, Med Access, Practice Solutions Suite, Healthquest, Telin) used by the primary care practice (p<0.001). Of the five EMR systems included, it was observed that Med Access contributed the largest proportion of patients with missing or undefined smoking data (n=13,110; 66.4%). When examining the missingness of smoking records by each EMR system, Practice Solutions Suite had the most complete (97% of these patients had a smoking record) and Healthquest had the least complete (23% of these patients had a smoking record); however, the total number of patients in these groups were small (462 patients on Practice Solutions EMR; 644 patients on Healthquest EMR).

Lastly, it was also observed that patients who were missing smoking status also had more missing data for other clinical variables (BMI and blood pressure).

4.4.1 Methods 1 and 2 (Multiple Imputation)

Table 4.2 summarizes patient smoking statuses for those without missing data (complete cases), MI for all three statuses (current, past, never; Method 1) and MI for only past or never smokers (Method 2). Both methods resulted in complete data for all patients in the full cohort; however, the proportion of current smokers varied substantially, with 25.3% in Method 1 and 12.5% in Method 2. Compared to the Alberta-specific CTADS survey data,⁹⁷ Method 1 may be overestimating the proportion of current smokers and Method 2 may be underestimating current smokers. The ‘past’ category for Methods 1 and 2 falls within the survey’s estimated 95% confidence intervals. The relative error for the multiple imputation, which compared the proportions of smoking status in the complete cases to the imputed dataset, was small (current=-0.05%, past=0.12%, never=-0.03%). Any additional free text related to smoking status was deemed to be insufficient for the MI process, as this was rarely available for most patients and was highly unstructured and not easily interpretable (e.g., over 850 unique strings were found with an average of 18 characters within each string; examples of text entries included “smokes socially”, “smoker: quit”, “up to 2ppd”, “yes”, “tobacco use assessment”, etc.).

4.4.2 Method 3 (Pattern-Matching Algorithm)

To validate the accuracy of the pattern-matching algorithm on all available smoking records, all 3295 unique pairs of original Smoking Status and Smoking Value data were manually reviewed from a total of 110,850 smoking records for 39,898 patients. The pattern-matching algorithm correctly assigned a calculated smoking status for 103,935 records (93.8%) (Table 4.3). In the data review, two new categories were included: ‘Conflicting Information’, which reported on pairs of Status and Value data elements in the same smoking record that were

contradicting (e.g. record containing Status = “Current” and Value = “non smoker”); and ‘Not Smoking-related’, which identified records that had been categorized as a smoking record but contained non-smoking information (e.g., blister pack medications). The smoking categories that were most accurately coded by the algorithm were ‘past’ (99.9%), ‘never’ (99.3%), ‘not current’ (90.4%), and ‘current’ (83.2%), while the ‘unknown’ category had the least accurate classification rate at 63%. The low accuracy in ‘unknown’ classification results largely because many of the records contain multiple risk factors (such as alcohol use, diet, exercise, etc.). Since the smoking classification algorithm is part of a general risk factor classification algorithm, it is designed to produce a definitive classification when only a single risk factor is identified in a given record.

When examining the most recent smoking status for each patient calculated by the pattern-matching algorithm, the proportion of current smokers was slightly lower than in the complete case data (18.2% vs. 21.2%) but was similar to the survey prevalence of 18.9%. Although the algorithm may have more accurately recoded the smoking data, 23.6% of patients were still missing smoking status; however, this is a reduction by 17 percentage points in missingness as compared to the raw, uncoded data of the complete cases.

4.5 Discussion

This study examined post-extraction methods for improving the quality of smoking data in a primary care EMR database. The methodologies explored here differed by their aims and processes – the goal of MI is to improve the completeness of data using other variables in the dataset while also accounting for variability of these data. While the fully conditional multiple imputation successfully eliminated all missing smoking data, there is uncertainty around the

accuracy of the smoking status classification. We used a province-specific general population survey to estimate the external validity of the smoking classifications produced by the MI; however, the CTADS survey data have their own limitations, such as a low response rate, small sample size, and non-response bias,⁹⁸ which means the true prevalence of smoking is still unclear and particularly for adults with hypertension. Thus, the smoking information produced by MI is more appropriate for epidemiological purposes rather than for use in clinical feedback or practice quality improvement.

The rationale for smoking status as MNAR data in the MI was that current smokers are more likely to be captured in the EMR data, given its importance as a cardiovascular risk factor. Method 2 (MNAR) resulted in a lower proportion of current smokers (12.5%) compared to the other methods and survey data, which ranged from 18.2-25.3% (Table 4.2). It may be possible that for older adults with chronic conditions, emphasis is placed on smoking cessation and therefore, lower proportions of current smokers are observed in this cohort compared to population-level surveys. This finding is similar to a Canadian direct health measures survey, where only 9.9% of females and 14.7% of males treated for hypertension were current smokers; this analysis was focused on a smaller sample of older adults aged 60-79 (n=2,111).⁹⁹ Furthermore, the proportion of smokers was found to decline with increasing age in a U.K. primary care database, which was also accompanied by a rising proportion of former smokers as age increased.

The pattern-matching algorithm was designed to use additional information in the smoking record to produce more complete and accurate smoking statuses on an individual level, which is the recommended method for purposes where the accuracy of an individual's smoking status has greater significance – for example, when using primary care EMR data for practice

quality improvement, clinical feedback, or generating a cohort or registry of current smokers. Although this method resulted in reasonably accurate smoking status classifications, the algorithm was still not able to classify 24% of smoking records; further, the category of ‘not current’ is ambiguous and may not be useful information on its own, as this may represent either former or never smokers (though, the number of total records in this category was small). These issues underscore the larger challenges associated with unstructured, inconsistent textual data in primary care EMRs. Any coding or classification that is applied to the CPCSSN data is predicated on the data being present and accurate in the record. The source of missing or inaccurate data can be difficult to identify, as there are numerous places throughout the data flow pathway that can contribute to poor data quality.⁷⁶ This includes during point-of-care encounters (patient self-reporting to clinician; poor data entry; lack of structured fields in the EMR to enter smoking information or EMR containing multiple fields where conflicting information is entered), during data extraction, or through data processing (misclassification or errors in the coding algorithm).^{76,83}

Ultimately, a variety of approaches are required to achieve complete, consistent, and accurate information about patient smoking status in EMRs. Any post-extraction approach to data quality improvement relies on complete and accurate data, which therefore means that strategies to support meaningful use of EMRs at the point of care are critical. For example, we found a statistically significant difference in the recording of smoking status attributed to type of EMR system. There are many different EMR products that exist with no formal requirements on the type or location of fields included in the EMR or how the interface appears to the user. Some systems allow for more structured, consistent information to be recorded in an easy-to-navigate way, though some contain more unstructured free text or check-boxes in hard-to-access areas of

the chart that can be more difficult for the extraction and interpretation of data. In addition to developing better tools for providers, there have been other examples of interventions to enhance data quality at the practice-level: employing a dedicated data entry clerk was found to improve the quality and usability of EMR data (particularly for smoking status), though the long-term sustainability and scale may not be feasible;⁴³ implementing data quality feedback tools, reports, and educational training for clinicians have been shown to improve the recording of information in electronic health records,^{41,42} as well as reduce variations in quality due to different electronic systems;⁴² and lastly, policy-based initiatives, such as meaningful use regulations and financial incentives, have had a positive impact on the recording of smoking status, among other data elements.^{100,101} Even so, data improvement at the point-of-care require more intensive interventions and resources, and therefore post-data extraction strategies should also be pursued. This includes further exploration of the methods described in this paper, as well as more advanced techniques like natural language processing.⁴⁵

4.5.1 Limitations

The first limitation of this study is the lack of a true reference standard from which to assess external validity. By not knowing the true prevalence of smokers (and other smoking statuses) among those with hypertension, we cannot say for certain whether the MI methods are correct from a population-level. The CTADS survey included in the paper reported on the smoking status of individuals in the general Alberta population and was not specific to hypertension, unlike the patient sample here. Other sources were considered but not used for various reasons: the Canadian Community Health Survey (CCHS) and the Canadian Health Measures Survey (CHMS) do not have publicly available smoking rates for people with

hypertension. The CCHS describes the proportion of self-reported current (daily or occasional) smokers but not former or non-smokers. The CHMS only includes participants up to the age of 79.

The CPCSSN primary care EMR data also have specific shortcomings. Most importantly, the smoking data within the EMR may be limited in terms of its accuracy and timeliness, which would subsequently impact the results of the MI and pattern matching algorithm. Since smoking status is a dynamic state that can change at any point in time, it is difficult to confirm whether the data present in the EMR are a true reflection of each patient's current smoking status. We attempted to address this by selecting the most recent smoking status available in the EMR for each patient and found that the majority of this cohort had their most recent smoking status recorded within the previous two years.

The CPCSSN data represent a relatively small sample of providers (just over 5% of all family physicians in Alberta⁷³) and patients in Alberta (from a provincial population of over 4.3 million in 2019⁸⁵), and thus may be reasonably but not entirely representative of all providers and patients in the province or elsewhere in Canada. Both the MI and pattern-matching algorithms may perform differently depending on variations between different regional databases and the EMR system from which the data are extracted (e.g., only 5 out of the 11 EMR systems within the CPCSSN database were represented here).

4.6 Conclusion

Multiple imputation and algorithmic pattern-matching are two ways that EMR data can be improved post-extraction, with MI returning complete data and the classification algorithm permitting accuracy estimates. There is potential to further enhance the algorithm using more

complex coding with more data points, which would further improve its accuracy and ability to reduce missing patient smoking status. In the future, these methods could also be tested on other risk factor information in a primary care EMR databases, such as alcohol use, diet and exercise.

Table 4.1 Demographic and clinical characteristics of patients with and without smoking status recorded in their EMR

Variable	Smoking Status Recorded (N=28,634)	Missing Smoking Status (N=19,743)	p-value
Female, n (%)	14,547 (50.8)	10,342 (52.4)	0.001
Missing sex, n (%)	1 (0.0)	1 (0.0)	--
Age in years, mean (SD)	64.9 (13.4)	65.3 (14.7)	0.001
BMI in kg/m ² , mean (SD)	30.9 (6.8)	30.7 (6.8)	0.001
Missing BMI, n (%)	1660 (5.8)	3334 (16.9)	--
Systolic BP, mean (SD)	132.3 (16.2)	133.4 (16.5)	<0.001
Diastolic BP, mean (SD)	78.9 (10.8)	79.2 (10.7)	0.027
Missing BP, n (%)	84 (0.3)	406 (2.0)	--
COPD, n (%)	3078 (10.7)	1645 (8.3)	<0.001
Diabetes, n (%)	8040 (28.1)	4920 (24.9)	<0.001
Type of EMR system, n (%)			<0.001
Wolf	20,471 (71.5)	5872 (29.7)	
Med Access	7481 (26.1)	13,110 (66.4)	
Practice Solutions Suite	448 (1.6)	14 (0.1)	
Telin	86 (0.3)	251 (1.3)	
Healthquest	148 (0.5)	496 (2.5)	
Urban residence, n (%)	20,373 (72.6)	15,728 (81.1)	<0.001
Missing postal code, n (%)	558 (1.9)	344 (1.7)	--
Number of total clinical encounters, median (lower quartile, upper quartile)	37.0 (20, 63)	24.0 (12, 44)	<0.001

Abbreviations: BMI=body mass index; BP=blood pressure; COPD=chronic obstructive pulmonary disease; EMR=electronic medical record; SD=standard deviation

Table 4.2 Comparison of different improvement methods for smoking categories

	2017 CTADS General Population Survey (Alberta) N=1444	CPCSSN Hypertensive Patients (Alberta) N=48,377			
Category	Survey Prevalence % (95% CI)	Complete Cases n (%)	Method 1: MAR Imputation* n (%)	Method 2: MNAR Imputation** n (%)	Method 3: CPCSSN Pattern Matching n (%)
Current smoker	18.9 (14.8-22.9)	6061 (21.2)	12,255 (25.3)	6061 (12.5)	6729 (18.2)
Past/former smoker	21.1 (17.0-25.1)	6494 (22.7)	9628 (19.9)	11,305 (23.4)	8868 (24.0)
Never smoker	60.1 (55.2-65.0)	16,079 (56.2)	26,494 (54.8)	31,011 (64.1)	20,846 (56.4)
Not current	--	--	--	--	502 (1.4)
Missing, n/N (%)	--	19,743/48,377 (40.8)	0/48,377 (0)	0/48,377 (0)	11,432/48,377 (23.6)

Abbreviations: CI=confidence interval; CPCSSN=Canadian Primary Care Sentinel Surveillance Network; CTADS=Canadian Tobacco, Alcohol and Drug Survey

*MAR=Missing at Random; assumes missing values are random for all smoking statuses (current, past, never)

**MNAR=Missing Not at Random; assumes all current smokers have been documented and missing values are random for never or past smokers

Table 4.3 Comparison of the pattern-matching algorithm classification to data review

		Data Review (reference standard)							
		Current	Never	Not Current	Past	Unknown	Conflicting Information	Not Smoking-related	Accuracy (%)
Pattern-Matching Algorithm	Current	23,858	5	243	298	1562	2,688	28	83.2
	Never	0	53,211	32	49	1	274	0	99.3
	Not Current	0	0	2,696	2	284	0	0	90.4
	Past	2	1	1	21,691	2	6	0	99.9
	Unknown	236	13	819	282	2,479	20	67	63.3

Figure 4.1 Flow diagram for the pattern-matching smoking status classifier (Method 3)

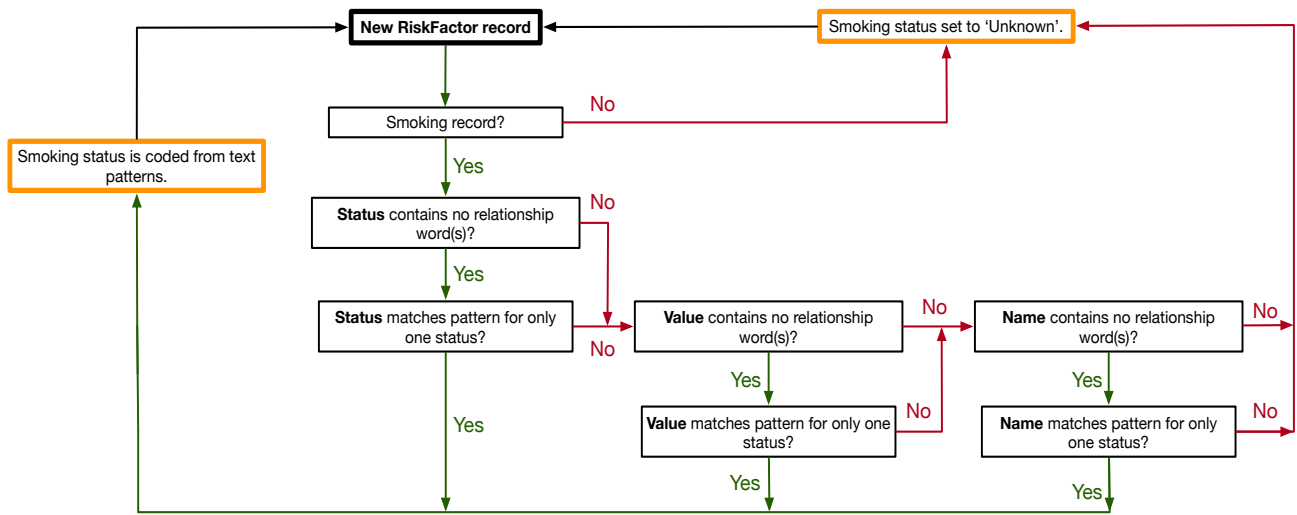
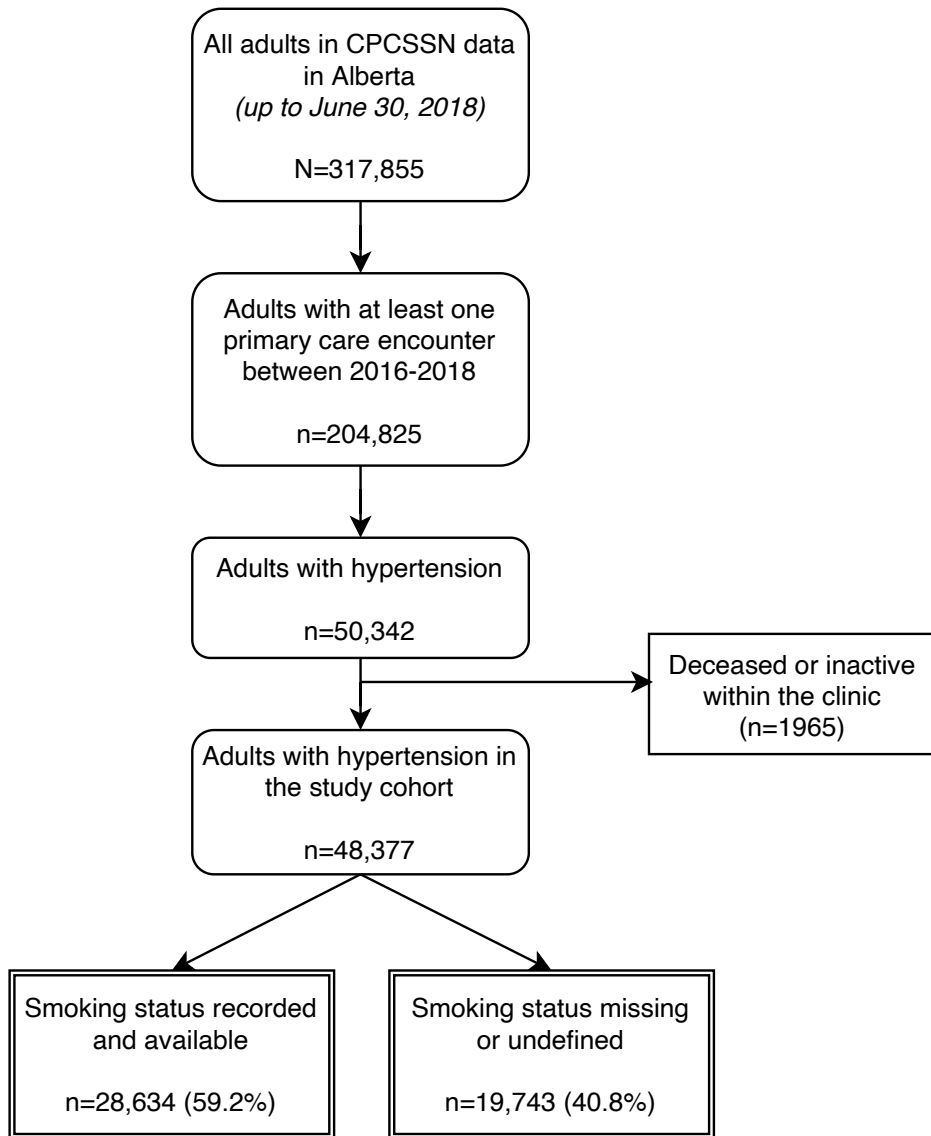


Figure 4.2 Flow diagram for patients in the cohort study



**CHAPTER FIVE: PRIMARY CARE EMR AND ADMINISTRATIVE DATA
LINKAGE IN ALBERTA, CANADA: DESCRIBING THE SUITABILITY
FOR HYPERTENSION SURVEILLANCE**

This paper is under revision at *BMJ Health & Care Informatics*:

Garies S, Youngson E, Soos B, Forst B, Duerksen K, Manca D, McBrien K, Drummond N, Quan H, Williamson T. Primary care EMR and administrative data linkage in Alberta, Canada: describing the suitability for hypertension surveillance. *BMJ Health & Care Informatics* (in revision, June 2020).

5.1 Abstract

Objective: To describe the process for linking electronic medical record (EMR) and administrative data in Alberta and examine the advantages and limitations of utilizing linked data for hypertension surveillance.

Methods: De-identified EMR data from 323 primary care providers contributing to the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) in Alberta were used. Mapping files from each contributing provider were generated from their EMR to facilitate linkage to administrative data within the provincial health data warehouse. Deterministic linkage was conducted using valid personal healthcare number (PHN) with age and/or sex. Characteristics of patients and providers in the linked cohort were compared to population-level sources. Criteria used to define hypertension in both sources were examined.

Results: Data were successfully linked for 6,307 hypertensive patients (96.2% of eligible patients) from 49 contributing providers. Non-linkages from invalid PHN (n=246) occurred more for deceased patients and those with fewer primary care encounters, with differences due to type of EMR and patient EMR status. The linked cohort had more patients who were female, >60 years, and residing rurally compared to the provincial healthcare registry. Family physicians were more often female and medically trained in Canada compared to all physicians in Alberta. Most patients (>97%) had ≥ 1 record in the registry, pharmacy, emergency/ambulatory care, and claims databases; 44.3% had ≥ 1 record in the hospital discharge database.

Conclusion: EMR-administrative data linkage has the potential to enhance hypertension surveillance. The current linkage process in Alberta is limited and subject to selection bias. Processes to address these deficiencies are underway.

5.2 Background

Chronic disease surveillance is an important public health function, which includes monitoring health events over time and examining processes of disease such as etiology, treatment patterns, long-term management, and health outcomes.¹⁰²⁻¹⁰⁴ In order to be effective, a robust surveillance system should include information that is current and timely, comprehensive, accurate, accessible at various levels, longitudinal, stable, flexible, cost-effective, and population-based or representative of the population.^{102,105,106}

In Canada, administrative data are often used for chronic disease surveillance, as it is routinely-collected, longitudinal information on healthcare encounters for nearly the entire population.^{86,104} The proliferation of electronic medical record (EMR) systems in healthcare settings has now established EMR data as another option for disease surveillance.^{9,107} Both sources are not without limitations: administrative data lack important clinical indicators and risk factors, while EMR data are extracted for a small subset of the population and are more difficult to analyze.

Linking EMR and administrative data may help to alleviate respective shortcomings and enhance surveillance systems, yet this does not occur routinely in Canada for a variety of reasons: linkages must be conducted within each province and territory (if they are done at all) due to distinct provincial/territorial policies and health information legislation. Linkage processes can also vary depending on differences in health information laws, data holdings, and data accessibility for secondary uses.

We used primary care EMR data from one Canadian province (Alberta) to conduct linkages with five administrative databases for patients with hypertension. Hypertension was chosen as an example chronic condition, as it has been identified as an important surveillance

priority and is largely managed in primary care.^{6,108} Our objective was to describe the current process for EMR-administrative data linkage in Alberta and to examine the advantages, potential biases, and limitations of using linked data for hypertension surveillance.

5.3 Methods

5.3.1 Data Sources

5.3.1.1 Primary Care EMR Data

The Canadian Primary Care Sentinel Surveillance Network (CPCSSN) is a national repository of de-identified primary care EMR data for nearly 2 million patients, which is available for approved secondary purposes such as research and surveillance.^{17,72} CPCSSN is a collaboration of practice-based research networks in seven Canadian provinces. Each regional network extracts, processes, and standardizes EMR data from participating family physicians and nurse practitioners (all termed “sentinels”), which is then submitted to the national repository. The CPCSSN data include patient demographics, diagnoses, prescribed medications, physical measurements (e.g., blood pressure, height, weight), physician billing claims, behavioural risk factors, laboratory results, referrals, and medical procedures. The national CPCSSN database includes approximately 4.8% of all Canadians as of December 2019⁵⁸ and has been found to be slightly overrepresented by females and older adults, though this is consistent with primary care populations.³¹ Family physicians who contribute data to CPCSSN constitute an estimated 3.3% of all family physicians in Canada.¹⁰⁹

Alberta, one of 13 Canadian provinces and territories, is located in the western part of the country with a population of just under 4.4 million people in 2019.⁵⁸ In Alberta, the Northern and Southern Alberta Primary Care Research Networks (NAPCRen and SAPCRen, respectively)

extract de-identified EMR data twice annually from five commonly-used EMR systems. Longitudinal patient data are available dating back to the start of EMR use, which varies by clinic and by patient, until the most recent data extraction (up to June 30, 2018 for this study). Included in this sample were data for 204,825 adult patients and 310 primary care providers, representing 4.8% and 5.6% of the total Alberta citizens and family physicians, respectively. Patients who have explicitly opted out of the CPCSSN database are excluded. There are more females, older adults, and rural patients in the CPCSSN database for Alberta as compared to the provincial healthcare registry, which is typically used as a population denominator (Appendix 1). Additional details about the CPCSSN data extraction and processing in Alberta have been published previously.⁸³

5.3.1.2 Administrative Data

Alberta Health Services (AHS) is the provincial health authority and has legislative permission to hold identifiable patient-level administrative data. Five administrative databases were used and linked through the AHS Analytics Enterprise Data Warehouse:

1. Discharge Abstract Database (DAD) – in-patient hospital discharges from all acute care institutions in Alberta. Each record contains the most responsible diagnosis and up to 24 secondary diagnosis fields coded using the International Classification of Diseases version 10, Canadian Enhancement (ICD-10-CA).
2. National Ambulatory Care Reporting System (NACRS) – ambulatory care visits and procedures (e.g., day surgery, emergency room visits, community rehabilitation services). NACRS contains a primary diagnosis with up to nine secondary codes for each record using ICD-10-CA.

3. Practitioner claims – submitted billing claims from fee-for-service physicians and other allied health practitioners, in addition to shadow-billed claims from alternative payment models, with information about the type of provider and service provided.
4. Pharmaceutical Information Network (PIN) – dispensed prescription medications from approximately 99% of community pharmacists in Alberta. Details include drug name/identification number, dose, quantity dispensed, and supply amount. PIN does not capture medications dispensed in hospitals or emergency departments.
5. Alberta Health Care Insurance Plan (AHCIP) registry – population-level demographic registry of all Albertans registered for publicly-funded healthcare in the province; includes nearly every resident except members of Canadian Armed Forces and inmates of federal penitentiaries (who are instead covered by the federal government).

5.3.2 Data Linkage

Figure 5.1 outlines the current process for linking de-identified EMR data from CPCSSN with identifiable administrative data from AHS. CPCSSN sentinels were required to formally agree to the data linkage through a research agreement specific to this project. Sentinels who agreed to the data linkage were asked to generate an EMR Mapping File from their clinic EMR that included patients' personal healthcare number (PHN), EMR ID, sex, and date of birth. PHN is a unique identifier assigned to residents of Alberta who have registered with the AHCIP. Additionally, CPCSSN generated a 'CPCSSN' Mapping File that contained two variables for each hypertensive patient in the CPCSSN database belonging to a participating sentinel: EMR ID (collected at time of CPCSSN's routine data extraction) and CPCSSN ID (a unique, randomly-

assigned study ID). The EMR and CPCSSN Mapping Files were securely transferred to AHS Analytics.

Within AHS Analytics, the EMR Mapping Files from each clinic were deterministically linked to the administrative data using the patient PHN, as well as sex and date of birth for additional verification. The CPCSSN ID was then added to the administrative data using the EMR ID, which is an identifier that is common to both mapping files. Direct patient identifiers were removed prior to secure transfer of the data to the University of Calgary. The de-identified CPCSSN data were transferred separately from CPCSSN to the University of Calgary, where it was linked to the administrative data using CPCSSN ID.

5.3.3 Defining Hypertension

Patients with hypertension were identified using validated case definitions for both data sources (Box 1). The administrative definition is based on 2 physician claims using relevant ICD-9 codes within 2 years or 1 hospitalization using ICD-10 codes within the DAD.¹¹⁰ The EMR-based definition was developed by CPCSSN and uses a combination of text words, ICD-9 diagnostic codes, and antihypertensive medications located throughout the EMR; when validated, this definition performed slightly better than the administrative definition.¹⁸

5.3.4 Analysis

The overall linkage rate was calculated as the proportion of eligible hypertensive patients in the CPCSSN cohort that matched exactly by PHN (and one of sex or date of birth) to the AHCIP registry. Patient and clinic characteristics for those who were in the linked cohort and

those who could not be linked were compared using a chi-squared test for categorical data, t-test for continuous variables, and Wilcoxon test for non-parametric data.

Selected demographic characteristics of adult patients in the linked hypertensive cohort were reported and compared to characteristics for all hypertensive patients in the full CPCSSN data for Alberta. Characteristics of family physicians contributing linked data were compared to all physicians in the CPCSSN database in Alberta and to all family medicine physicians in Alberta according to the Canadian Institute for Health Information (CIHI) *Supply, Distribution and Migration of Physicians in Canada* report.¹⁰⁹

The number of patients meeting any of the criteria from the administrative and EMR definitions for hypertension was explored by sex and age group. All data analysis was performed in RStudio version 1.1.456.

5.3.5 Ethics Approval

As part of the CPCSSN project, both NAPCReN and SAPCReN have research ethics approval to extract de-identified EMR data through their respective universities. The linkage study was granted separate approval by the Conjoint Health Research Ethics Board at the University of Calgary and the Health Research Ethics Board at the University of Alberta.

5.4 Results

5.4.1 Data Linkage

Figure 5.2 presents the flow of patients throughout the linkage process. Within the CPCSSN database, 50,342 adult patients were identified as having hypertension (24.5%). Fifty-five sentinels out of a possible 243 CPCSSN sentinels across Alberta had explicitly agreed to the

data linkage, however, we were unable to link data for 6 sentinel (two were excluded due to data extraction issues with the EMR vendor; four providers were unable to have their EMR Mapping File submitted from the clinic). From the remaining 49 sentinels, a total of 6,553 patients were eligible for linkage with administrative data.

Overall, we were able to link 6,307 of the 6,553 (96.2%) eligible patients from the CPCSSN EMR cohort to the administrative data. The proportion of patients in the CPCSSN data who had at least one record in each of the administrative databases was very high (100% for AHCIP Registry, PIN, and Claims; 97.2% for NACRS) or expected (44.3% for DAD in-patient hospitalizations). Of those whose data were unsuccessfully linked, the majority (n=200; 81.3%) was due to having no available PHN; the remaining 18.7% (n=46) of unlinked patients had mismatched sex and/or date of birth between the EMR Mapping File and AHCIP Registry.

Table 5.1 describes selected characteristics of patients who had their EMR data linked to administrative data and those whose data could not be linked. Patients with unlinked CPCSSN records were more likely to be deceased according to the EMR (24.4% compared to 0.5% in the linked cohort; $p<0.001$) and were also associated with a lower mean number of primary care encounters (39.2 compared to 45.8 in the linked cohort; $p=0.013$). Patients with linked records also differed by EMR status ($p<0.001$) and the type of EMR system used at the practice ($p<0.001$) compared to those with unlinked records.

5.4.2 Patient Cohort Comparisons

Table 5.2 describes the demographic characteristics of adult patients with hypertension in the linked cohort (n=6,307) compared to all adults with hypertension in the full CPCSSN dataset from Alberta (n=50,342). The linked cohort had slightly more females (53.8% vs. 51.4%),

similar mean ages (64.7 vs. 65.4 years), and fewer urban residents (69.5% vs. 76.1%) than all patients with hypertension in the full CPCSSN database for Alberta. The full CPCSSN database compared to the healthcare registry (population denominator) was observed to be more overrepresented by females (56.3%), older patients within several age bands, and rural patients (19.2% vs. 9.0%) (Appendix A).

5.4.3 Provider Comparisons

Table 5.3 describes the characteristics of family physicians who completed the data linkage (n=48) compared to two groups: 1) all family medicine physicians in the province; and 2) family physician sentinels contributing data to CPCSSN dataset within Alberta. The linked cohort included more physicians who were female (56.2% vs. 42.7%), fewer who were internationally-trained (2.1% vs. 43.2%), and slightly less from rural practices (10.4% vs. 12.2%) when compared to all family physicians in Alberta.¹⁰⁹ Although the mean age was similar, the linked cohort had higher proportions of physicians between the ages of 30-39 years and 60-69 years. Compared to all participating CPCSSN sentinels in Alberta, family physicians contributing to the linkage were observed to be slightly younger (mean age 48.7 years in linked cohort versus 47.1) with lower proportions of internationally-trained physicians and those practicing in rural locations.

5.4.4 Hypertension Definitions

Overall, there were 1,276 patients (20.2%) that met all 5 criteria from both the administrative and EMR definitions; nearly 9% of patients (n=564) met at least one component of the EMR hypertension definition, but did not meet any administrative criteria (data not

shown). From the two administrative criteria, most patients met the billing claims criterion; from the three EMR criteria, the EMR profile was the most frequently met criterion (n=4,801), though the billing claims (n=4,693) and prescribed medication (n=4,449) also had similar numbers of patients (Figure 5.3). Major discrepancies between definitions by sex and age group were not evident.

5.5 Discussion

This paper describes the process for linking de-identified primary care EMR data with administrative data in Alberta. Combining these sources increased the breadth of hypertension-related data, particularly with the addition of longitudinal blood pressure values and other physical measurements to several administrative data sources. However, the number of patients in the final linked cohort was small and the current process for data linkage may not be practical for province-level surveillance. Our focus has been on augmenting existing surveillance activities, but linked administrative and primary care EMR data could be a valuable addition for use in health services research, retrospective cohort studies, and practice quality improvement.

5.5.1 Advantages

5.5.1.1 Comprehensiveness

The most significant advantage of this type of data linkage is the creation of more comprehensive information about people with hypertension, more so than either data source on its own. For instance, smoking status, weight trajectories over time, medication prescriptions and longitudinal blood pressure measurements recorded in primary care practices are now available alongside pharmacy-dispensed medication records and acute outcomes requiring hospitalization,

which does not exist currently in Alberta. This combination of community-based, detailed clinical information and system-wide healthcare encounters can help to enhance our understanding of disease processes, management and outcomes.

5.5.1.2 Completeness

Linking EMR and administrative data can also help to improve data completeness and evaluate the accuracy of data elements. The Electronic Medical Record Administrative data Linked Database (EMRALD) in Ontario, Canada has demonstrated this in their assessment of prescribed medication validity in EMRs compared to a provincial drug database, as well as using linked data to confirm a variety of administrative-based disease identification algorithms.^{50,111–113}

5.5.1.3 Timeliness

Both administrative and EMR data are available in a timeframe suitable for chronic disease surveillance, as real-time data are not necessarily warranted for conditions that have a long latency period. However, the linkage process described here took just over one year to complete, after research ethics approvals, obtaining sentinel agreements to link EMR data, generating the mapping files at each of the participating clinics, and completing the EMR and administrative linkage. Even so, updating linkages once per year, for example, would still be reasonable for hypertension or other chronic disease surveillance.

5.5.2 Limitation and Potential Biases

5.5.2.1 Coverage of Capture

The primary care setting represents citizens who have accessed the first level of the healthcare system and the CPCSSN database is a further subset of these patients whose providers consented to participate. The CPCSSN data from Alberta were observed to include higher proportions of physicians who were female, completed medical training in Canada, and practice in rural locations. Patients in the Alberta CPCSSN data were more likely to be female, older, and rurally-located compared to all AHCIP registrants, although this reflects a typical primary care population in terms of older adults and women.³¹

One major limitation of this current approach to EMR-administrative data linkage is the potential for selection bias to occur. Data required for surveillance should ideally be population-based or at least representative of the broader population. There was a clear discrepancy in the low proportion of urban patients included in the linked dataset; in addition, the providers contributing to the CPCSSN database and the linkage may potentially differ from the broader population of providers for reasons we are unable to measure (e.g. more comfortable with data or research; record information differently in the EMR).

5.5.2.2 Completeness & Accuracy

Although this linkage brought together previously discrete data about patients with hypertension, risk factors and other information relevant to hypertension surveillance are still missing or incomplete. For example, smoking status and alcohol use are often entered inconsistently as free text in EMRs and may not be recorded for all patients in an extractable format, which limits their useability.^{33,80} Environmental and socioeconomic factors also

contribute to the etiology and management of hypertension; however, neither primary care EMR nor administrative health data capture this information sufficiently. This highlights the need for additional data linkage to occur with multiple health and non-health sources (e.g. ministerial data from occupation, education, social services; surveys; GIS mapping), while still respecting individual privacy and maintaining minimal risk of re-identification.

There were a few issues with the accuracy of the linkage process. Duplicate patients were found in the AHCIP Registry due to different year of birth or sex recorded between subsequent registration years; however, this represented a very small proportion of patients (0.1%). Another challenge related to the process of generating the mapping files. Because CPCSSN extracts data from five different EMR systems in the province, there are variations in the EMR data extraction methods (which also generates the CPCSSN mapping files) and creation of the EMR mapping files from each clinic, as well as differences in how patients are assigned (or not assigned) to providers in the EMR. This was evident in the differing linkage rates by EMR type, by deceased status (which is not necessarily known or recorded in primary care settings) and in the EMR status of patients in the clinic (e.g. whether only active or all patients were included in the mapping files). The process for extracting the mapping files in the clinic EMR and the criteria for patient selection requires further standardization.

5.5.2.3 Misclassification

Misclassifying those who truly have or do not have hypertension by the case finding algorithms can greatly bias epidemiological or surveillance outcomes. Previous validation studies have shown that the EMR-based definition for hypertension outperformed the administrative definition in both sensitivity (84.5% vs. 75%) and PPV (92.9% vs. 81%).^{18,110} In

this cohort, 564 (8.9%) of the hypertensive patients in the EMR were not identified as hypertensive in the administrative data; this may be due to a longer timeframe of available data in the EMR or less frequent hypertension billing claims submitted for patients with long-term control.

Poor data quality can also contribute to misclassification through data entry errors in the clinic EMR from the use of non-specific or no diagnostic codes, information recorded in inaccessible formats (e.g. free text notes; scanned documents), or errors produced during CPCSSN processing.¹¹⁴ Since the administrative definition for hypertension only uses the ICD-9 billing codes and a smaller number of ICD-10-CA codes within in-patient data, this may further limit the ability of the administrative data to identify all true cases of hypertension and thus, linkage with detailed EMR data can augment case definition accuracy.

5.6 Conclusion

EMR-administrative linkage can result in novel, rich data for hypertension surveillance. The linkage process described here is limited in terms of its small sample size and generalizability; however, this work will inform the development of a more efficient and robust process for data linkage. In addition to surveillance, this type of linkage could also be used to strengthen retrospective cohort studies by supplementing administrative data with the rich details found in EMR data. Expanding linkages to include data from other sources, as well as increasing the number of practitioners contributing to the linkage, would further enhance the applicability for hypertension surveillance and epidemiology.

Table 5.1 Comparison of characteristics for hypertension patients with and without linked data

	Linked (n=6,307)	Not Linked (n=246)	p-value
Female, n (%)	3395 (53.8)	127 (51.6)	0.539
Age, mean (SD)	64.7 (14.1)	65.2 (17.2)	0.523
Deceased year recorded in the EMR, n (%)	30 (0.5)	60 (24.4)	<0.001
Patient EMR status			<0.001
Active, n (%)	5326 (84.4)	126 (51.2)	
Deceased, n (%)	104 (1.6)	60 (24.4)	
Inactive, n (%)	99 (1.6)	8 (3.3)	
Unknown, n (%)	778 (12.3)	52 (21.1)	
Urban residence, n (%)*	4312 (69.5)	156 (65.5)	0.225
Type of EMR			<0.001
Wolf, n (%)	3266 (51.8)	221 (89.8)	
Med Access, n (%)	3041 (48.2)	25 (10.2)	
Number of primary care encounters, mean (SD)	45.8 (40.7)	39.2 (47.2)	0.013

Abbreviations: EMR=electronic medical record; SD=standard deviation

*Postal code for determining urban or rural residence was missing for 99 patients (1.6%) in the Linked cohort and 8 patients (3.3%) in the Not Linked cohort

Table 5.2 Comparison of hypertensive patient characteristics in the CPCSSN database in Alberta and the linked EMR-administrative cohort

	Primary Care EMR Data (all patients with hypertension)	Linked Admin-EMR Data (hypertension cohort)
Data source	CPCSSN data in Alberta up to June 30, 2018; patients with at least 1 visit in last 2 years	CPCSSN data in Alberta linked to AHCIP Registry & other admin databases
Total adults, N	50,342	6,307
Female, n (%)	25,865 (51.4)	3,395 (53.8)
Male, n (%)	24,475 (48.6)	2,912 (46.2)
Age, mean (SD)	65.4 (14.1)	64.7 (14.1)
Age groups, n (%)		
20-39 years	2247 (4.5)	330 (5.2)
40-59 years	14,020 (27.8)	1790 (28.4)
60-69 years	14,030 (27.9)	1833 (29.1)
70-79 years	11,662 (23.2)	1397 (22.1)
80 years and older	8383 (16.7)	957 (15.2)
Urban, n (%)	37,507 (76.1)	4,312 (69.5)
Rural, n (%)	11,783 (23.9)	1,896 (30.5)
Missing/unknown residence or postal code, n (%)	1,053 (2.1)	99 (1.6)

Abbreviations: AHCIP=Alberta Health Care Insurance Plan; CPCSSN=Canadian Primary Care Sentinel Surveillance Network; EMR=electronic medical record; SD=standard deviation

Table 5.3 Comparison of family physician characteristics in Alberta and in the CPCSSN data

	All Family Medicine Physicians in Alberta¹⁰⁹	Family Physicians Contributing to CPCSSN AB^a	Family Physicians Who Agreed to EMR-Admin Data Linkage^b
Total, N	5,489	310	48*
Female, n (%)	2,343 (42.7)	171 (55.2)	27 (56.2)
Male, n (%)	3,146 (57.3)	139 (44.8)	21 (43.8)
Mean age, years (SD)	48.8	47.1 (10.3)	48.7 (12.0)
<30, n (%)	217 (4.0)	1 (0.3)	0 (0)
30-39, n (%)	1,354 (24.7)	67 (21.6)	13 (31.7)
40-49, n (%)	1,441 (26.3)	79 (25.5)	9 (22.0)
50-59, n (%)	1,191 (21.7)	72 (23.2)	7 (17.1)
60-69, n (%)	542 (9.9)	35 (11.3)	11 (26.8)
70+, n (%)	731 (13.3)	3 (1.0)	1 (2.4)
Missing age, n (%)	13 (0.2)	53 (17.1)	7 (14.6)
Rural practice, n (%)	670 (12.2)	43 (13.9)	5 (10.4)
International medical training, n (%)	2,373 (43.2)	69 (22.3)	1 (2.1)
Missing location of medical training, n (%)	n/a	19 (6.1)	1 (2.1)

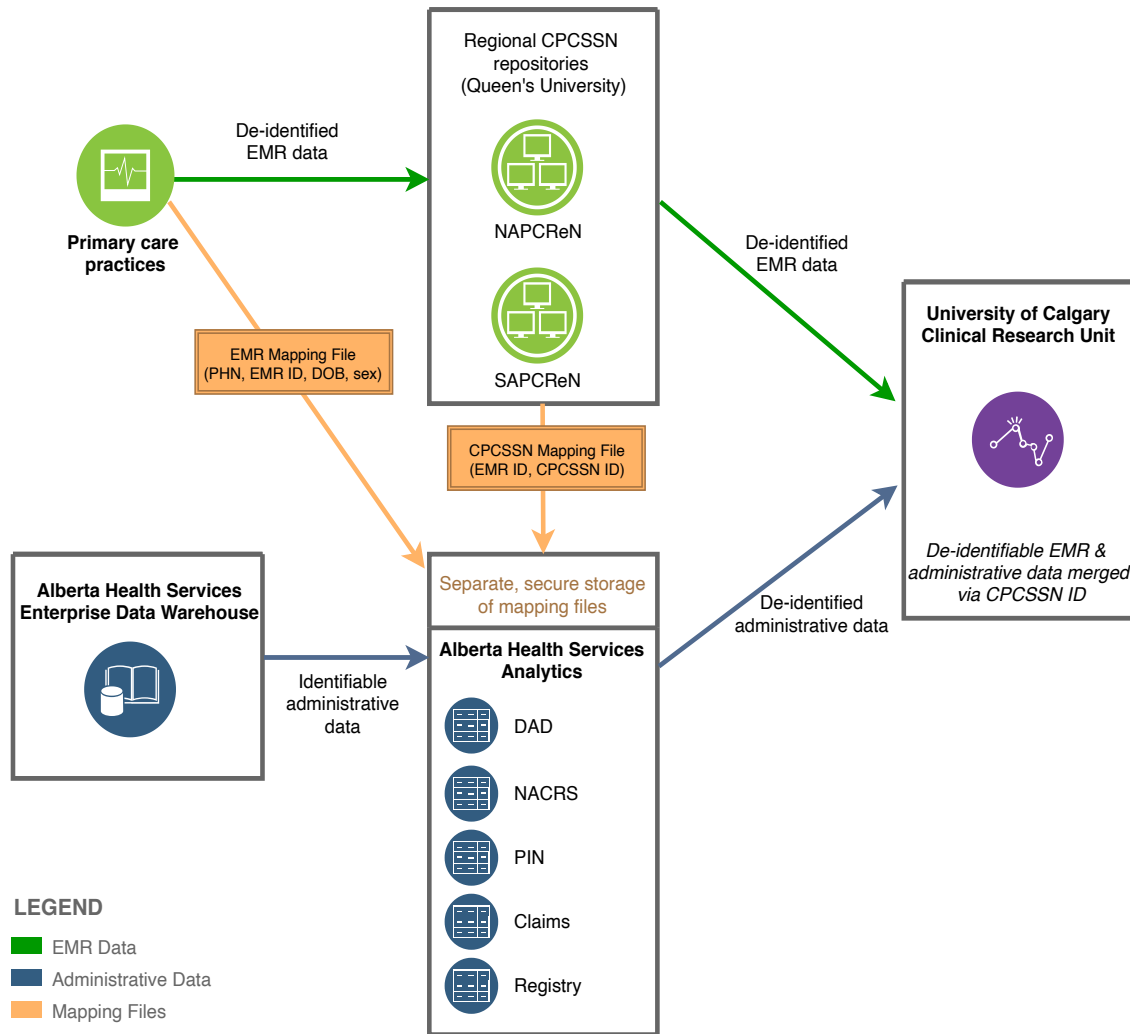
Abbreviations: AB=Alberta; CPCSSN=Canadian Primary Care Sentinel Surveillance Network; EMR=electronic medical record; SD=standard deviation

*Excludes one nurse practitioner from the 49 providers who contributed to the linkage

^a All data up to June 2018

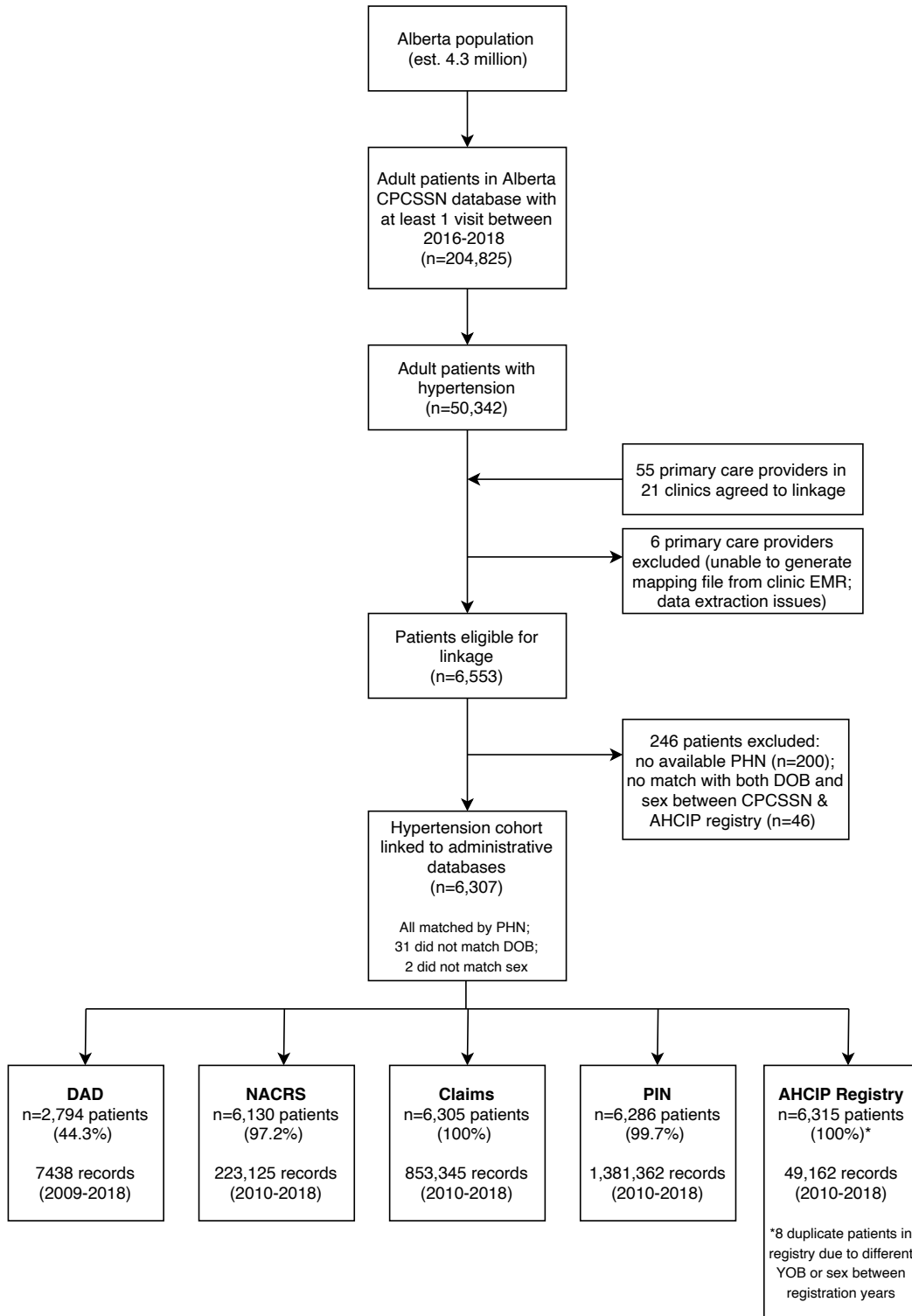
^b Data for hypertension cohort up to June 2018

Figure 5.1 Data flow and linkage process for primary care EMR and administrative data in Alberta



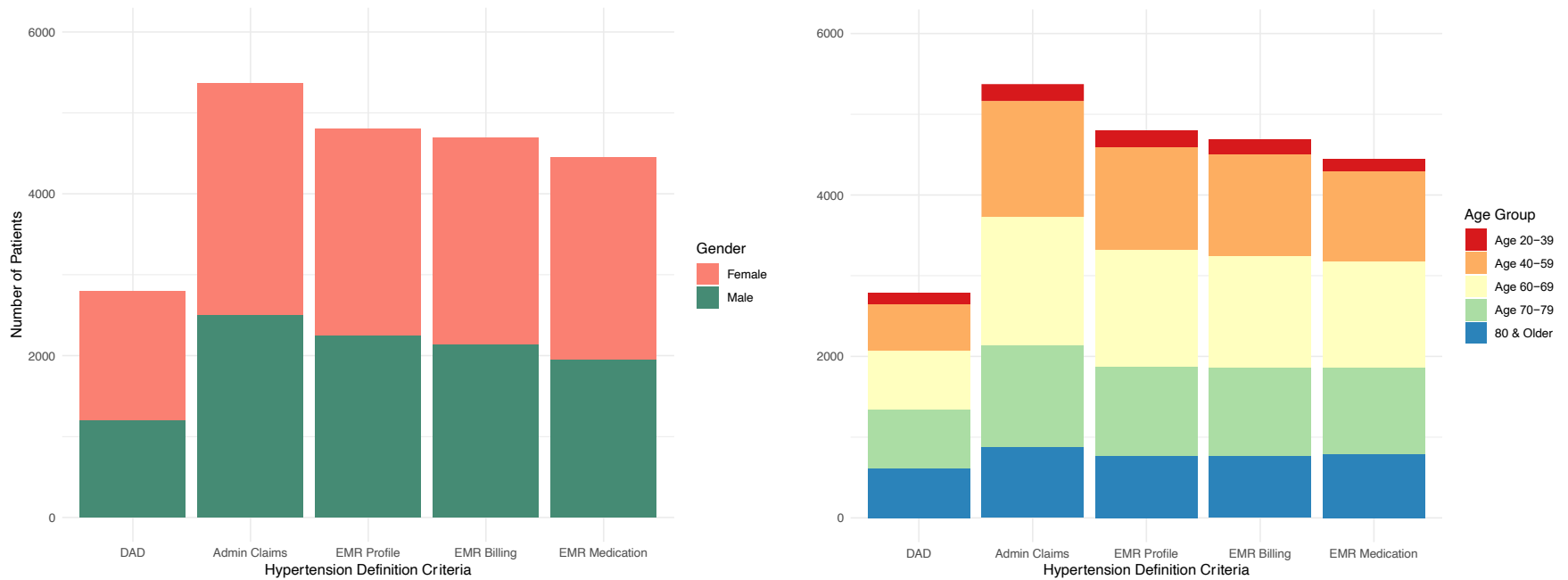
Abbreviations: CPCSSN=Canadian Primary Care Sentinel Surveillance Network; DAD=Discharge Abstract Database; DOB=date of birth; EMR=electronic medical record; NACRS=National Ambulatory Care Reporting System; NAPCReN=Northern Alberta Primary Care Research Network; PHN=personal healthcare number; PIN=Pharmaceutical Information Network; SAPCReN=Southern Alberta Primary Care Research Network

Figure 5.2 Flow diagram of patient selection into the linked hypertension cohort



Abbreviations: AHCIP=Alberta Health Care Insurance Registry; CPCSSN=Canadian Primary Care Sentinel Surveillance Network; DAD=Discharge Abstract Database; DOB=date of birth; EMR=electronic medical record; NACRS=National Ambulatory Care Reporting System; PHN=personal healthcare number; PIN=Pharmaceutical Information Network; YOB=year of birth

Figure 5.3 Patients in the linked data cohort meeting hypertension case criteria for administrative and EMR definitions by sex and age group (n=6,307)



Box 5.1 Criteria for hypertension definitions in administrative and EMR data

Definition	Data Source	Criteria	Codes	Validity
Administrative ¹¹⁰	Physician claims	At least two billing claims within two years	ICD9 codes: 401-405	Sensitivity: 75% Specificity: 94% PPV: 81% NPV: 92%
	Discharge Abstract Database (DAD)	One in-patient diagnosis code at any time	ICD10 codes: I10-I15	
EMR / CPCSSN ¹⁸	Physician Claims table	At least two billing claims within two years	ICD9 codes: 401-405	Sensitivity: 84.9% Specificity: 93.5 PPV: 92.9% NPV: 86.0%
	Problem List / Profile table	At least one diagnosis code at any time	ICD9 codes: 401-405	
	Prescribed Medication table	Any occurrence of a specified hypertension medication, <i>except</i> if one of the following co-morbid conditions exists: diabetes mellitus (250), tremor (333.1), migraine (346), myocardial infarction (410, 412), angina or cardiac dysrhythmias (413, 427), heart failure (428), esophageal varices (456.0, 456.1), calculus of kidney and ureter (592), portal hypertension (572.3)	ATC codes: C02, C03AA03, C03BA04, C03BA08, C03BA11, C03DB01, C03DB02, C03EA01, C07AA06, C07AB03, C07AB04, C07AG02, C07CB03, C08CA01, C08CA02, C08DA01, C09AA01, C09AA02, C09AA03, C09AA07, C09AA08, C09AA09, C09AA10, C09BA02, C09BA03, C09CA02, C09CA03, C09CA04, C09CA07, C09DA01, C09DA02, C09DA04, C09XA02	

Abbreviations: ATC=Anatomical Therapeutic Chemical (classification); CPCSSN=Canadian Primary Care Sentinel Surveillance Network; EMR=electronic medical record; ICD=International Classification of Diseases; NPV=negative predictive value; PPV=positive predictive value

Box 5.2 Criteria for a robust chronic disease surveillance system

System Features ^{102,105,106}
<ul style="list-style-type: none">▪ Current; timely▪ Comprehensive▪ Accurate▪ Accessible at various levels▪ Longitudinal▪ Stable▪ Flexible▪ Cost-effective▪ Population-based or representative of the population
Information Needs for Hypertension Surveillance
<ul style="list-style-type: none">▪ Biological risk factors (e.g. blood pressure, obesity/body mass index, pre-existing conditions, cholesterol & other lab markers)▪ Lifestyle risk factors (e.g. smoking status, physical inactivity, alcohol use, diet)▪ Disease progression (e.g. lab results, severity of disease)▪ Disease management (e.g. medications prescribed and dispensed; medical procedures)▪ Health outcomes (e.g. subsequent diagnoses, acute events requiring hospitalization, mortality)▪ Environmental and policy factors▪ Sociodemographics (e.g. ethnicity, occupation, poverty)▪ Patient-reported outcome measures and experiences (e.g. symptom burden, functional status, quality of life)▪ Healthcare costs

CHAPTER SIX: SUMMARY

6.1 Summary of Key Findings

Primary care EMR data are relatively new compared to administrative data, yet they contain a wealth of detailed patient-level information that can enhance current chronic disease surveillance systems and other secondary uses. However, it is critical to understand how EMR data are generated, extracted, and processed to inform where potential quality issues and biases may be occurring. My thesis work fills a gap in our knowledge about the lifecycle of CPCSSN data in Alberta and describes the content of the CPCSSN data in detail within a hypertension context. From this, an area of poor data quality was identified (smoking status), and two methods were tested to explore whether data completeness and accuracy could be improved by post-extraction analyses. Finally, EMR data for a cohort of patients with hypertension were linked to five administrative databases; this process was described in detail and the resulting linked data were examined for suitability for hypertension surveillance.

6.1.1 Describing Data Quality Within a Hypertension Context

Producing publicly available documentation describing CPCSSN data capture, extraction, and processing is critical for data users, but documentation is often a neglected aspect of data production. The documentation presented here examines the full lifecycle of CPCSSN data in Alberta and details all post-extraction transformations that take place. This work will improve our understanding of the data and, more importantly, will help to identify potential data issues, biases, and limitations. This information is imperative, particularly for a pan-Canadian database like CPCSSN, as each regional network can vary in their local data extraction and transformation processes, as well as in provincial/territorial health information policies, physician payment models, delivery of care, and types of EMR systems available.

The results of the data quality assessment show variability in the completeness and plausibility of selected data elements, as well as observed variation due to type of EMR system and between clinic sites. Blood pressure and antihypertensive medications are relatively complete and appeared to contain minimal outliers. Other relevant data elements are inconsistent in terms of completeness and plausibility; this includes height, weight, smoking, alcohol use, and ethnicity. These data would not be readily useable without additional post-extraction processing (e.g. pattern-matching or natural language processing for ethnicity or risk factor text data) and careful consideration of analytic choices (e.g. complete case analysis; using mean/median for multiple physical exam values; timeframe for exam value inclusion). Linkages to other data sources or improvements made to data entry at each site will further improve data quality, but are more challenging and require long-term strategies.

Despite these imperfections, primary care EMR data make valuable contributions to hypertension surveillance – blood pressure measurements in particular are a distinct advantage over administrative data-based surveillance systems (which contain no blood pressure or physical measurements) and surveys (which sometimes includes blood pressure measurements, but not longitudinally). The limitations of CPCSSN data should be considered before any secondary use and managed appropriately; this may include age-sex standardization, given the non-population-based sample; assessing representativeness for generalizing to a broader population; and considering potential misclassification of the condition(s) of interest.

6.1.2 Post-Extraction Methods for Improving EMR Data Quality

After conducting the data quality assessment, I focused on statistical methods to improve the quality of smoking data. In the CPCSSN data from Alberta, only 59.2% of hypertensive

patients had information on smoking status, despite this being an important cardiovascular risk factor.⁹⁰ I tested two post-extraction methods for improving data quality, which differed by their aims, processes, and outcomes:

1. Fully conditional multiple imputation (MI) was used to improve the completeness of smoking data by replacing missing smoking statuses with imputed values. MI is advantageous over other methods (e.g. complete case analysis) because it produces imputed values based on the distribution of other variables in the dataset, which accounts for variability. Unlike complete case analyses, MI also retains the entire sample and therefore does not lose statistical power. In this study, MI was successful in producing complete smoking information for all patients; however, it was difficult to quantify the accuracy of the MI results. Estimated relative errors were small when testing MI in a complete case dataset with smoking values removed in proportion to the 'real' dataset. Even so, there was no appropriate external reference that could be used to compare proportions of smoking status categories derived from the MI. It was also not possible to confirm whether the smoking information in this cohort were missing at random, missing not at random, or missing completely at random. Even though MI is considered a more flexible model for handling different types of missingness, errors can occur if not all variables associated with the missing data are specified, especially for data that are missing not at random.⁹² Based on this, MI would not be suitable for purposes that required high accuracy at the individual level, such as quality improvement or certain research applications. If a comparable external source was available, MI could potentially be used to fill in missing data gaps for more broad epidemiological purposes.

2. A pattern-matching algorithm was developed to improve both completeness and accuracy of patient smoking status by incorporating free text from other areas of the smoking record and classifying each patient's status as current, never, past, not current, or unknown. Although the pattern-matching algorithm was not able to categorize 23.6% of patients, this method is recommended based on a balance of improvements in completeness and known levels of accuracy (~94%). However, further refinements need to be made to incorporate more complexity into the algorithm before widespread application to the CPCSSN data.

MI and pattern-matching can be used to improve some aspects of data quality in smoking records, though the application depends on the intended use of the data. Regardless, there is a limit to the success of pre-extraction processing and analyses in the improvement of data quality if the data are not reasonably complete or accurate from the source EMR. There is evidence to suggest that the completeness of EMR data may improve over time as a function of the length of time a physician has been using the clinic EMR.⁸² However, it is likely this would only reach a limited threshold, given the number of additional factors influencing data quality at the practice level, such as the types and location of fields available through the EMR interface, data entry preferences, perceived value of good quality data, knowledge around use of EMR systems, and coding, clinician time, and practice workload.⁷⁶

6.1.3 Linking EMR Data with Administrative Data

It was anticipated that linking population-level administrative data with detailed clinical data from EMRs would enhance disease surveillance systems, with the assumption that 'more data are better'. I linked CPCSSN EMR and administrative data for patients with hypertension,

which resulted in a small cohort with linked data (n=6,307). Because each CPCSSN sentinel is required to provide explicit permission for the linkage of their de-identified EMR data, the final sample was not completely representative of providers in the larger CPCSSN database or the broader population. Those who contributed data for the linkage were more often female physicians, a larger proportion of physicians that were medically trained within Canada, and slightly fewer physicians practicing in rural areas, compared to all family physicians in Alberta. The linked cohort is a sample taken from a sample of patients and providers within the CPCSSN database, which is derived from the larger population of citizens in Alberta. The CPCSSN data in Alberta itself is overrepresented by females, older patients, and rural residents.

Although the current linkage process is not yet ideal for contributing to province-wide surveillance activities, this linked hypertension cohort may be useful for health outcomes research and validation studies, with the recognition of limitations related to selection bias and generalizability. There are several novel uses of these linked data that could be explored: comparing medications prescribed in primary care and medications dispensed in a pharmacy; using longitudinal patient-level blood pressure and weight measurements to validate diagnostic codes or case definitions related to hypertension and obesity in administrative data; understanding disease trajectories or healthcare utilization throughout various settings while adjusting for or stratifying by risk factors (e.g. weight/BMI, smoking).

In the pursuit of achieving high quality, ‘big’ linked data for disease surveillance, this work highlighted several challenges that need to be addressed:

1. Both the process for generating mapping files from each clinic EMR and the criteria for which patients are included in these files needs to be standardized and more consistent. This was the source of discrepancies in the quality of the linkage, where

- various types of EMR systems produced different mapping files, and some clinics included/excluded patients in the mapping files by various statuses (e.g. all, active only, no deceased/inactive).
2. Creating these EMR mapping files should ideally be conducted at a regular time point and stored in a central, secure location for use in subsequent data linkage projects. This would allow the data linkage to be conducted more efficiently and predictably.
 3. Sentinel agreement for this linkage work was low and resulted in small sample size. This is not surprising, given the busy nature of clinical work and potential uneasiness with EMR-administrative data linkage, which is not routinely conducted in Alberta. Efforts are needed to socialize the benefits of data linkage for clinicians and to obtain larger numbers of agreement for subsequent linkage requests, which includes expanding the broader CPCSSN project to include routine data linkage and the inclusion of linked EMR-administrative data in the CPCSSN data holdings.

6.2 Implications for EMR Data Users

In a 2017 report, the Auditor General of Alberta lamented on the challenges of healthcare integration in Alberta, in part due to the “lack of sharing and use of clinical information”.¹¹⁵ These issues may be considerably improved through the implementation of Connect Care, a new province-wide electronic health system utilizing Epic software.¹¹⁶ This integration of over 1,300 clinical information systems in Alberta has already begun; however, Connect Care will not include most family physicians except for a small number of practices that operate within AHS facilities. Although the provincial Community Information Integration (CII) initiative has been designed to address this failing by incorporating a summary of primary care patients’ EMR

information into Epic, this is not easily executed and only contains a limited snapshot of the patient EMR. Accordingly, the ongoing work by CPCSSN to liberate valuable clinical information from primary care EMRs plays an ever-important role in ensuring evidence-informed decision-making through the use of high-quality data for health systems improvement and better patient care. As both CPCSSN and Connect Care continue to expand, there will be an increasing number of diverse EMR data users; this may include clinicians, researchers, epidemiologists, analysts, decision-makers, and health systems/organizations. Thus, the findings from this work will have different implications depending on the intended use and context.

Overall, this work highlights the significant amount of post-extraction processing that is necessary to transform raw EMR data into useable information. The detailed documentation of these processes is critical for data users and should be adopted and replicated by other CPCSSN PBRNs to describe their own methods. Further, resolving deficiencies in EMR data quality will require a multifaceted approach and should involve a variety of settings and stakeholders, including clinicians, EMR vendors, and CPCSSN data managers.

For researchers or analysts intending to use the CPCSSN data for hypertension surveillance or research, this work provides information about the CPCSSN data in Alberta that can help users determine whether the data are appropriate for their needs and to ensure any limitations are known so that appropriate inferences can be made about subsequent findings. This type of transparent, comprehensive documentation is not widely or easily available, even though it is desirable for the initial planning stages of research studies or surveillance activities, as well as necessary for assisting analysts disentangle the complex CPCSSN database. This work also demonstrates that EMR-administrative linkage is possible, although it is likely not ready for large-scale disease surveillance until the linkage processes become more efficient and less

biased. Notwithstanding, there is great potential on the horizon for the emergence of a novel, rich linked data source to enhance current surveillance activities and health outcomes research.

There are also implications for clinical users of EMR data. Data quality issues can potentially impact patient care or distort outcomes from practice improvement initiatives if the data entered in the EMR are not complete, accurate, and current. Evaluating EMR data quality is unlikely to be at the forefront of clinical priorities within busy practices, and therefore CPCSSN should play a role in facilitating data quality reporting for clinics and physicians in an effective and actionable way.

6.3 Strengths and Limitations

6.3.1 Strengths

EMR data are the future of ‘big health data’ and central to innovative health applications, such as learning health systems and personalized medicine.^{76,116} The CPCSSN database is unique in Canada and the cleaned, coded, accessible primary care EMR data for over 300,000 patients in Alberta does not exist anywhere else in the province. As data quality plays a fundamental role as a precursor for all secondary uses, these findings are novel and can be applied to CPCSSN data in other provinces/territories or differing contexts. The data capture, extraction, and processing documentation and data quality assessment demonstrate the advantages of the CPCSSN data as compared to the original source EMR, which includes many automated processes for cleaning raw, messy data and mapping values or text to standard classification systems. This work benefitted immensely as a result of the high-level of technical capacity and knowledge of the CPCSSN team in Alberta, which was indispensable for the construction of the detailed data quality documentation in Chapter 2. In addition, the EMR data quality assessment and reporting

used an existing framework specific to distributed data networks, such as CPCSSN, which also expanded on the well-established STROBE guidelines.^{3,118}

Alberta also hosts an extensive set of provincial administrative databases that have been made more accessible and timelier largely due to substantial investments through the Alberta Strategies for Patient Oriented Research (SPOR) SUPPORT (Support for People and Patient-Oriented Research and Trials) Unit Data Platform, which made this linkage work feasible within a reasonable timeframe.

This work has a specific focus in primary care, which represents the majority of healthcare encounters and is the setting for most chronic disease diagnosis, treatment, and management.¹¹⁹ The advantages of working in this context is that the application of research can benefit a broader section of the healthcare system, patients, and providers, as opposed to a more narrow scope focused on specialist care or hospital utilization.¹¹⁹

6.3.2 Limitations

There are a number of limitations to address. The description of the CPCSSN capture, extraction, and processing only includes minimal information about how data are entered or gathered at the point of care. This varies greatly by provider, clinic, and EMR system; however, it was not possible to describe or quantify this within the scope of the current project. Secondly, the data quality assessment in Chapter 3 may not have encompassed all data elements that some would deem important for hypertension surveillance, such as cholesterol values, diet, or physical activity. Furthermore, the data quality assessment focused on single element descriptive statistics and many aspects of quality were not able to be addressed, such as accuracy (as compared to a reference standard), temporal constraints (expected changes over time), and consistency/cross-

validation with other data elements or databases.³ The data quality assessment was also not stratified by EMR system to a great extent, which is often a significant source of variations in data quality. Thus, identifying the source of quality issues (e.g. no field available in EMR, error in data entry, patients not being asked about risk behaviours, issues/errors with CPCSSN processes) may still be challenging and requires further work.

On several occasions, a true reference standard from which to assess accuracy or external validity was not available or inadequate. EMR chart reviews are often used as the reference standard; however, they can be difficult to access and for many structured data fields (e.g. height, weight), errors or omissions in the CPCSSN data would also exist in the source EMR. Physician SOAP notes and PDF documents (e.g. referral/specialist letters, diagnostic imaging) in the EMR are currently not extracted by CPCSSN, though they may contain additional information and context to augment and verify the accuracy and completeness of CPCSSN data. There are important ethical, patient privacy, and data security considerations, as well as the application of advanced analytics, that are required before it would be possible to extract and utilize the rich narrative text from these parts of the EMR.

With respect to the linkage of EMR and administrative data, the current process described here is cumbersome and requires improvement. Fifty-five out of 243 (22.6%) CPCSSN-participating primary care providers explicitly agreed to the linkage work, though it was only possible to link data for 49 providers in total (20.2%). This resulted in a small number of patients in the linked cohort (n=6,307), which was likely not suitable for surveillance purposes. The sample of providers contributing to the linkage was also small and non-representative, potentially resulting in selection and/or non-participation bias.

Lastly, this work represents only one province in Canada and the findings and/or linkage processes may not be generalizable to CPCSSN networks in other provinces or territories, particularly among those networks extracting from EMR systems that are different from the five that were included in the Alberta extractions. However, the data quality assessment and code can be used as a template to be replicated and modified accordingly across multiple CPCSSN networks in various jurisdictions.

6.4 Future Directions

6.4.1 Web-Based Data Quality Reporting

There has been renewed interest in evaluating and improving data quality throughout CPCSSN nationally, as well as a long-standing need for more transparent, readily available documentation about CPCSSN data and its quality. The work discussed here will form the basis for developing a web-based data quality dashboard intended to inform prospective researchers and data users about the quality of CPCSSN data within various contexts and information needs. This can be featured on the national CPCSSN website (cpcssn.ca) and/or run on regional networks' data and hosted on local websites. The dashboard will feature all available CPCSSN data elements and will provide basic statistics (e.g. range, mean/median, frequency), with a dynamic function that will allow users to display results for selected elements for specific patient and disease contexts. The technical expertise and web software have been established for previous CPCSSN dashboard development (<http://alberta.cpcssn.ca/the-dpt>) and can be similarly implemented for the data quality reporting.

A descriptive data quality assessment will also be provided for CPCSSN sentinel clinicians, who currently receive online feedback reports after each data extraction cycle (twice

yearly). Sentinels receive their CPCSSN reports through a web-based link that provides them with a summary of demographic and clinical characteristics for their patient panel with comparisons at various levels (e.g. practice, Primary Care Network, provincial). These are produced by CPCSSN regional data managers, who have the technical skills to integrate additional metrics into the existing reports. The data quality assessment for sentinels will be brief but would focus on data elements that are important for a clinical context, such as proportion of hypertensive patients missing blood pressure (within a specific time frame); proportion of patients with diabetes who do not have this diagnosis recorded in the patient profile/problem list; proportion of adults without a smoking status, etc. Data quality feedback reports such as this have been produced in other similar primary care EMR networks (e.g. The Health Information Network) and have been shown to improve the quality of information recorded in primary care EMRs.^{41,42} The impact of clinician data quality reports on the improvement of CPCSSN data could be evaluated using a time series design, where select data quality indicators are measured at regular intervals both before and after distributing sentinel data quality reports. This study design would be relatively simple to implement but would not capture data quality improvements due to other reasons (e.g. data entry initiatives within the clinic; changes to CPCSSN extraction or processing).

6.4.2 Additional Data Quality Improvements

This work explored two methods for enhancing data quality post-extraction: multiple imputation to address missingness and a pattern-matching algorithm to improve both completeness and accuracy. These methods can also be tested on other EMR data elements (e.g. alcohol use, diet, weight, height), depending on agreed priorities within the CPCSSN database,

which largely derives from recommendations made by the CPCSSN Data Quality Working Group. The pattern-matching algorithm described in Chapter 4 will be applied to the national CPCSSN dataset and further refinements made to accommodate variations due to the additional EMR systems, regions, and providers that were not included in the Alberta sample. The eventual goal is to accurately categorize smoking status to provide this coded information to researchers accessing the CPCSSN database.

There are also a variety of other methods that can be used for post-extraction data quality issues. One burgeoning approach that is being tested by several other CPCSSN regional networks is the use of natural language processing (NLP) to extract useable information from the rich free-text SOAP notes in the EMR. This technique could be considered in the future for Alberta. However, it would require additional research ethics approvals and agreement from CPCSSN sentinels to allow the extraction and mining of their EMR notes. NLP is also more time-intensive than pattern-matching and should be weighed against the value and quantity of information that can be reliably derived from NLP.

6.4.3 Future Areas of Research

There are many new and innovative research opportunities that can be pursued with access to linked EMR-administrative health data. Using the existing linked dataset generated for my thesis, future research studies will focus on the quality and improvement of the linked data and linkage processes. First, specific contents of the EMR data will be verified using the administrative data; for example, the accuracy and completeness of referral information in the EMR can be confirmed using practitioner billing claims and NACRS, and the completeness of

acute health outcomes (e.g. stroke, myocardial infarction) recorded in the EMR data can be verified with hospitalization and emergency department information from NACRS and DAD.

Secondly, a probabilistic linkage process for combining de-identified databases will be explored using the linked EMR-administrative cohort as the reference standard. This work will seek to develop a method for using non-identifying patient-level elements (such as postal code, age, sex, health conditions) common to both the EMR and administrative databases to estimate the probability of successful matching. This method would achieve more robust linkages that overcome issues related to selection bias and generalizability; however, the accuracy of correct matches and non-matches may be lower or more uncertain, and this compromise should be evaluated appropriately within the context of intended use.

Finally, the healthcare utilization and health management patterns of patients with hypertension will be explored using the linked data. The novel combination of EMR and administrative data would allow for investigations into system utilization at various levels (e.g. primary care encounters, emergency department use, hospitalizations) and proportion of antihypertensive medications prescribed and dispensed (or not dispensed), all of which can be stratified by blood pressure levels (i.e. at target or not at target). Using physical measurements, like blood pressure, can help estimate disease severity and provides a better understanding of the patient context, as this is not currently possible using administrative data such as practitioner billing claims.

6.5 Conclusion

Primary care EMR data is a contemporary source of detailed patient information from community practices, though it requires substantial processing to clean, code, and standardize the

data into a more useable dataset. The data are suitable for hypertension surveillance and provide valuable, high-quality information about longitudinal blood pressure, prescribed antihypertensive medication, and diagnoses. Several data elements (e.g. ethnicity, weight, smoking) demonstrate poorer quality, and further improvements in their completeness and accuracy are necessary. Linking EMR data with administrative databases can provide a more robust, comprehensive source of data for disease surveillance and health outcomes research. The current process undertaken for linkage produced a small-sized cohort that is likely biased, which limits the generalizability and applicability for surveillance purposes. However, this work has informed the development of more efficient and effective processes for EMR-administrative linkages.

REFERENCES

1. Canadian Medical Association. National Results by FP/GP or Other Specialist, Gender, Age, and Province/Territory [Internet]. CMA Physician Workforce Survey 2017. 2017 [cited 2019 Jul 15]. Available from: <https://surveys.cma.ca/en/permalink/survey31>
2. Zelmer J, Hagens S. Advancing primary care use of electronic medical records in Canada. *Heal Reform Obs.* 2014;2(3):Article 2.
3. Kahn MG, Brown JS, Chun AT, Davidson BN, Meeker D, Ryan PB, et al. Transparent reporting of data quality in distributed data networks. *EGEMS (Washington, DC).* 2015;3(1):Article 7.
4. Quan H, McAlister FA, Khan N. The many faces of hypertension in Canada. *Curr Opin Cardiol.* 2014;29(4):354–9.
5. Padwal RS, Bienek A, McAlister FA, Campbell NR. Epidemiology of hypertension in Canada: an update. *Can J Cardiol.* 2016;32(5):687–94.
6. Godwin M, Williamson T, Khan S, Kaczorowski J, Asghari S, Morkem R, et al. Prevalence and management of hypertension in primary care practices with electronic medical records: a report from the Canadian Primary Care Sentinel Surveillance Network. *CMAJ Open.* 2015;3(1):E76–82.
7. Chang F, Gupta N. Progress in electronic medical record adoption in Canada. *Can Fam Physician.* 2015;61(12):1076–84.
8. Canada Health Infoway. Physicians' use of digital health and information technologies in practice [Internet]. 2018 Canadian Physician Survey. 2018 [cited 2019 Feb 28]. Available from: <https://www.infoway-inforoute.ca/en/component/edocman/3643-2018-canadian-physician-survey/view-document?Itemid=0>

9. Birtwhistle R, Williamson T. Primary care electronic medical records: A new data source for research in Canada. *CMAJ*. 2015;187(4):239–40.
10. van Staa TP, Goldacre B, Gulliford M, Cassell J, Pirmohamed M, Taweel A, et al. Pragmatic randomised trials using routine electronic health records: putting them to the test. *BMJ*. 2012;344:e55.
11. Gentil M-L, Cuggia M, Fiquet L, Hagenbourger C, Le Berre T, Banâtre A, et al. Factors influencing the development of primary care data collection projects from electronic health records: a systematic review of the literature. *BMC Med Inform Decis Mak*. 2017;17:139.
12. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, et al. The eMERGE Network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med Genomics*. 2011;4:1–11.
13. Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, Staa T van, et al. Data resource profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol*. 2015;44(3):827–36.
14. Medicines & Healthcare products Regulatory Agency; National Institute for Health Research. Clinical Practice Research Datalink [Internet]. 2020 [cited 2020 Jun 17]. Available from: <https://www.cprd.com>
15. Wolf A, Dedman D, Campbell J, Booth H, Lunn D, Chapman J, et al. Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. *Int J Epidemiol*. 2019;48(6):1740-1740g.
16. Netherlands Institute Health for Services Research. Nivel [Internet]. 2020. [cited 2020 Jun 17]. Available from: <https://www.nivel.nl/en/nivel-primary-care-database>

17. Garies S, Birtwhistle R, Drummond N, Queenan J, Williamson T. Data Resource Profile: National electronic medical record data from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN). *Int J Epidemiol.* 2017;46(4):1091-1092f.
18. Williamson T, Green ME, Birtwhistle R, Khan S, Garies S, Wong ST, et al. Validating the 8 CPCSSN case definitions for chronic disease surveillance in a primary care database of electronic health records. *Ann Fam Med.* 2014;12(4):367–72.
19. CPCSSN Team. Case definitions: Canadian Primary Care Sentinel Surveillance Network (CPCSSN), Version 2 [Internet]. 2019 [cited 2020 Mar 16]. Available from: <http://cpcssn.ca/wp-content/uploads/2019/07/CPCSSN-Case-Definitions-v2.pdf>
20. Greiver M, Williamson T, Barber D, Birtwhistle R, Aliarzadeh B, Khan S, et al. Prevalence and epidemiology of diabetes in Canadian Primary Care Practices: A report from the Canadian Primary Care Sentinel Surveillance Network. *Can J Diabetes.* 2014;38(3):179–85.
21. Williamson T, Natajaran N, O'Donnell DE, Khan S, Cave A, Green ME, et al. Chronic obstructive pulmonary disease in primary care: an epidemiologic cohort study from the Canadian Primary Care Sentinel Surveillance Network. *CMAJ Open.* 2015;3(1):E15–22.
22. Williamson T, Green ME, Jordan KP, Khan S, Birtwhistle R, Peat G, et al. Prevalence and management of osteoarthritis in primary care: an epidemiologic cohort study from the Canadian Primary Care Sentinel Surveillance Network. *CMAJ Open.* 2015;3(3):E270–5.
23. Williamson T, Khan S, Manca D, Birtwhistle R, Wong ST, Patten S, et al. The diagnosis of depression and its treatment in Canadian primary care practices: an epidemiological study. *CMAJ Open.* 2014;2(4):E337–42.

24. Khan S, Garies S, Drummond N, Molnar F, Birtwhistle R, Williamson T. Prevalence and management of dementia in primary care practices with electronic medical records: a report from the Canadian Primary Care Sentinel Surveillance Network. *CMAJ Open*. 2016;4(2):E177–84.
25. Rigobon A V, Birtwhistle R, Khan S, Barber D, Biro S, Morkem R, et al. Adult obesity prevalence in primary care users: An exploration using Canadian Primary Care Sentinel Surveillance Network (CPCSSN) data. *Can J Public Health*. 2015;106(5):e283–9.
26. Greiver M, Kalia S, Voruganti T, Aliarzadeh B, Moineddin R, Hinton W, et al. Trends in end digit preference for blood pressure and associations with cardiovascular outcomes in Canadian and UK primary care: A retrospective observational study. *BMJ Open*. 2019;9(1):1–11.
27. Garies S, Hao S, McBrien K, Williamson T, Peng M, Khan NA, et al. Prevalence of hypertension, treatment, and blood pressure targets in Canada associated with the 2017 American College of Cardiology and American Heart Association blood pressure guidelines. *JAMA Netw Open*. 2019;2(3):e190406.
28. Garies S, Irving A, Williamson T, Drummond N. Using EMR data to evaluate a physician-developed lifestyle plan for obese patients in primary care. *Can Fam Physician*. 2015;61(5):e225–31.
29. Birtwhistle R, Green ME, Frymire E, Dahrouge S, Whitehead M, Khan S, et al. Hospital admission rates and emergency department use in relation to glycated hemoglobin in people with diabetes mellitus: a linkage study using electronic medical record and administrative data in Ontario. *CMAJ Open*. 2017;5(3):E557–64.

30. Queenan JA, Birtwhistle R, Drummond N. Supporting primary care public health functions. *Can Fam Physician*. 2016;62(7):603–4.
31. Queenan JA, Williamson T, Khan S, Drummond N, Garies S, Morkem R, et al. Representativeness of patients and providers in the Canadian Primary Care Sentinel Surveillance Network: a cross-sectional study. *CMAJ Open*. 2016;4(1):e28–32.
32. Greiver M, Williamson T, Bennett TL, Drummond N, Savage C, Aliarzadeh B, et al. Developing a method to estimate practice denominators for a national Canadian electronic medical record database. *Fam Pract*. 2013;30(3):347–54.
33. Greiver M, Aliarzadeh B, Meaney C, Moineddin R, Southgate CA, Barber DTS, et al. Are we asking patients if they smoke?: Missing information on tobacco use in Canadian electronic medical records. *Am J Prev Med*. 2015;49(2):264–8.
34. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research. *J Am Med Inform Assoc*. 2013;20(1):144–51.
35. Casey JA, Schwartz BS, Stewart WF, Adler NE. Using electronic health records for population health research: a review of methods and applications. *Annu Rev Public Health*. 2016;37:61–81.
36. Wang R, Strong D. Beyond accuracy: what data quality means to data consumers. *J Manag Inf Syst*. 1996;12(4):5–33.
37. Thiru K, Hassey A, Sullivan F. Systematic review of scope and quality of electronic patient record data in primary care. *BMJ*. 2003;326:1070.
38. Bowen M, Lau F. Defining and evaluating electronic medical record data quality within the Canadian context. *Electron Healthc*. 2012;11(1):5–13.

39. Canadian Medical Association. How can Canada achieve enhanced use of electronic medical records? Toronto, ON; 2014.
40. Brouwer HJ, Bindels PJE, Van Weert HC. Data quality improvement in general practice. *Fam Pract*. 2006;23(5):529–36.
41. Taggart J, Liaw ST, Yu H. Structured data quality reports to improve EHR data quality. *Int J Med Inform*. 2015;84(12):1094–8.
42. van der Bij S, Khan N, Ten Veen P, de Bakker DH, Verheij RA, Blumenthal D, et al. Improving the quality of EHR recording in primary care: a data quality feedback tool. *J Am Med Inform Assoc*. 2016;356(24):2527–34.
43. Greiver M, Barnsley J, Aliarzadeh B, Krueger P, Moineddin R, Butt D, et al. Using a data entry clerk to improve data quality in primary care electronic medical records: a pilot study. *Inform Prim Care*. 2011;19(4):241–50.
44. McBrien KA, Souri S, Symonds NE, Rouhi A, Lethebe BC, Williamson TS, et al. Identification of validated case definitions for medical conditions used in primary care electronic medical record databases: a systematic review. *J Am Med Inform Assoc*. 2018;25(11):1567–78.
45. Liao KP, Cai T, Savova GK, Murphy SN, Karlson EW, Ananthakrishnan AN, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. *BMJ*. 2015;350:1–6.
46. Pearce C, McLeod A, Patrick J, Ferrigi J, Bainbridge MM, Rinehart N, et al. Coding and classifying GP data: The POLAR project. *BMJ Heal Care Informatics*. 2019;26:1–6.
47. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res*. 2004;32(Database issue):D267-270.

48. Wright A, Pang J, Feblowitz JC, Maloney FL, Wilcox AR, Ramelson HZ, et al. A method and knowledge base for automated inference of patient problems from structured data in an electronic medical record. *J Am Med Inform Assoc.* 2011;18:859–67.
49. Herrett E, Shah AD, Boggon R, Denaxas S, Smeeth L, Van Staa T, et al. Completeness and diagnostic validity of recording acute myocardial infarction events in primary care, hospital care, disease registry, and national mortality records: cohort study. *BMJ.* 2013;346(7909):1–12.
50. Schultz SE, Rothwell DM, Chen Z, Tu K. Identifying cases of congestive heart failure from administrative data: A validation study using primary care patient records. *Chronic Dis Inj Can.* 2013;33(3):160–6.
51. Tu K, Mitiku TF, Ivers NM, Guo H, Lu H, Jaakkimainen L, et al. Evaluation of Electronic Medical Record Administrative data Linked Database (EMRALD). *Am J Manag Care.* 2014;20(1):e15–21.
52. Canadian Institute for Health Information. Pan-Canadian primary health care electronic medical record content standard, version 3.0 [Internet]. Ottawa, Canada; 2014. Available from:
https://secure.cihi.ca/free_products/PHC_EMR_Content_Standard_V3.0_Business_View_EN.pdf
53. Wiebe N, Varela LO, Niven D, Ronksley PE, Iraggori N, Robertson HL, et al. Evaluation of interventions to improve inpatient hospital documentation within electronic health records: a systematic review. *Int J Popul Data Sci.* 2018;3(4):4–9.

54. Blak BT, Thompson M, Dattani H, Bourke A. Generalisability of The Health Improvement Network (THIN) database: demographics, chronic disease prevalence and mortality rates. *Inform Prim Care*. 2011;19(4):251–5.
55. Roland M, Guthrie B. Quality and Outcomes Framework: what have we learnt? *BMJ*. 2016;354:i4060.
56. Government of Canada. Canadian Chronic Disease Surveillance System: summary of methods [Internet]. 2019 [cited 2019 Dec 29]. Available from: <https://health-infobase.canada.ca/ccds/data-tool/Methods>
57. Statistics Canada. Canadian Community Health Survey - Annual Component (CCHS) [Internet]. 2020 [cited 2020 Mar 24]. Available from: <https://www.statcan.gc.ca/eng/survey/household/3226>
58. Statistics Canada. Population estimates, quarterly [Internet]. Table 17-10-0009-01. 2020 [cited 2020 Mar 24]. Available from: <https://doi.org/10.25318/1710000901-eng>
59. University of Manitoba. Manitoba Centre for Health Policy [Internet]. Manitoba Population Research Data Repository Data List. 2020 [cited 2020 Mar 29]. Available from: http://umanitoba.ca/faculties/health_sciences/medicine/units/community_health_sciences/departmental_units/mchp/resources/repository/datalist.html
60. Katz A, Enns J, Smith M, Burchill C, Turner K, Towns D. Population Data Centre Profile: The Manitoba Centre for Health Policy. *Int J Popul Data Sci*. 2020;4(2):1–10.
61. ICES. Institute for Clinical Evaluative Sciences [Internet]. 2020 [cited 2020 Mar 26]. Available from: <https://www.ices.on.ca>

62. University of Toronto. Family & Community Medicine [Internet]. UTOPIAN Data Safe Haven. 2020 [cited 2020 Mar 29]. Available from: <https://www.dfcm.utoronto.ca/utopian-data-safe-haven>
63. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G. Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. *Health Aff.* 2014;33(7):1123–31.
64. Kruse CS, Goswamy R, Raval Y, Marawi S. Challenges and opportunities of big data in health care: a systematic review. *JMIR Med Informatics.* 2016;4(4):e38.
65. Cave AJ, Davey C, Ahmadi E, Drummond N, Fuentes S, Kazemi-Bajestani SMR, et al. Development of a validated algorithm for the diagnosis of paediatric asthma in electronic medical records. *NPJ Prim Care Respir Med.* 2016;26:16085.
66. Lethebe BC, Williamson T, Garies S, McBrien K, Leduc C, Butalia S, et al. Developing a case definition for type 1 diabetes mellitus in a primary care electronic medical record database: an exploratory study. *CMAJ Open.* 2019;7(2):E246–51.
67. Queenan JA, Farahani P, Ehsani-Moghadam B, Birtwhistle R V. The prevalence and risk for herpes zoster infection in adult patients with diabetes mellitus in the Canadian Primary Care Sentinel Surveillance Network. *Can J Diabetes.* 2018;42(5):465–9.
68. Singer A, Yakubovich S, Kroeker AL, Dufault B, Duarte R, Katz A. Data quality of electronic medical records in Manitoba: Do problem lists accurately reflect chronic disease billing diagnoses? *J Am Med Inform Assoc.* 2016;23(6):1107–12.
69. Greiver M, Dahrouge S, O'Brien P, Manca D, Lussier MT, Wang J, et al. Improving care for elderly patients living with polypharmacy: Protocol for a pragmatic cluster randomized trial in community-based primary care practices in Canada. *Implement Sci.* 2019;14(1).

70. Reyes RR, Parker G, Garies S, Dolan C, Gerber S, Burton B, et al. Team-based comanagement of diabetes in rural primary care. *Can Fam Physician*. 2018;64(8):e346–53.
71. Brown F, Singer A, Katz A, Konrad G. Statin-prescribing trends for primary and secondary prevention of cardiovascular disease. *Can Fam Physician*. 2017;63(11):e495–503.
72. CPCSSN. Canadian Primary Care Sentinel Surveillance Network (CPCSSN) [Internet]. 2016 [cited 2019 Feb 14]. Available from: www.cpcssn.ca
73. Canadian Medical Association. Family medicine profile [Internet]. 2018. Available from: <https://www.cma.ca/sites/default/files/family-e.pdf>
74. Canadian Primary Care Sentinel Surveillance Network. CPCSSN Data Dictionary and ERD [Internet]. 2016 [cited 2019 May 23]. Available from: <http://cpcssn.ca/research-resources/cpcssn-data-dictionary-and-erd/>
75. Gamache P, Hamel D, Pampalon R. The material and social deprivation index: a summary [Internet]. 2017 [cited 2018 Nov 26]. Available from: <https://www.inspq.qc.ca/sites/default/files/santescope/indice-defavorisation/en/GuideMethodologiqueEN.pdf>
76. Verheij RA, Curcin V, Delaney BC, McGilchrist MM. Possible sources of bias in primary care electronic health record data use and reuse. *J Med Internet Res*. 2018;20(5):e185.
77. Murray M, Davies M, Boushon B. Panel size: how many patients can one doctor manage? *Fam Pract Manag*. 2007;14(4):44–51.
78. Public Health Agency of Canada. Report from the Canadian Chronic Disease Surveillance System: Hypertension in Canada, 2010. Ottawa, Canada; 2010.

79. Atwood KM, Robitaille CJ, Reimer K, Dai S, Johansen HL, Smith MJ. Comparison of diagnosed, self-reported, and physically-measured hypertension in Canada. *Can J Cardiol.* 2013;29(5):606–12.
80. Torti J, Duerksen K, Forst B, Salvalaggio G, Jackson D, Manca D. Documenting alcohol use in primary care in Alberta. *Can Fam Physician.* 2013;59(10):1128.
81. Terry AL, Stewart M, Cejic S, Marshall JN, de Lusignan S, Chesworth BM, et al. A basic model for assessing primary health care electronic medical record data quality. *BMC Med Inform Decis Mak.* 2019;19:30.
82. Tu K, Widdifield J, Young J, Oud W, Ivers NM, Butt DA, et al. Are family physicians comprehensively using electronic medical records such that the data can be used for secondary purposes? A Canadian perspective. *BMC Med Inform Decis Mak.* 2015;15(67).
83. Garies S, Cummings M, Forst B, McBrien K, Soos B, Taylor M, et al. Achieving quality primary care data: a description of the Canadian Primary Care Sentinel Surveillance Network data capture, extraction, and processing in Alberta. *Int J Popul Data Sci.* 2019;4(2).
84. Lewis S. A System in name only — access, variation, and reform in Canada’s provinces. *N Engl J Med.* 2015;372(6):497–500.
85. Government of Alberta. Quarterly population report; Second quarter 2019 [Internet]. 2019 [cited 2019 Oct 21]. Available from: <https://open.alberta.ca/dataset/aa3bce64-c5e6-4451-a4ac-cb2c58cb9d6b/resource/ae2c77eb-3ce0-4f70-90a3-6a2329f49355/download/2019-q2-population-report.pdf>

86. Public Health Agency of Canada. Public health infobase: Canadian Chronic Disease Surveillance System (CCDSS) [Internet]. CCDSS. 2017 [cited 2019 Mar 12]. Available from: <https://infobase.phac-aspc.gc.ca/ccdss-scsmc/data-tool/>
87. DeGuire J, Clarke J, Rouleau K, Roy J, Bushnik T. Blood pressure and hypertension. *Health Rep.* 2019;30(2):14–21.
88. Statistics Canada. Health fact sheets: chronic conditions, 2016 [Internet]. Ottawa, Canada; 2017. Available from: <https://www150.statcan.gc.ca/n1/en/pub/82-625-x/2017001/article/54858-eng.pdf?st=-6wP-Kde>
89. Vink NM, Klungel OH, Stolk RP, Denig P. Comparison of various measures for assessing medication refill adherence using prescription data. *Pharmacoepidemiol Drug Saf.* 2009;18:159–65.
90. Tobe SW, Stone JA, Anderson T, Bacon S, Cheng AY, Daskalopoulou SS, et al. Canadian Cardiovascular Harmonized National Guidelines Endeavour (C-CHANGE) guideline for the prevention and management of cardiovascular disease in primary care: 2018 update. *CMAJ.* 2018;190(40):E1192-206.
91. Lix L, Munakala SN, Singer A. Automated classification of alcohol use by text mining of electronic medical records. *Online J Public Health Inform.* 2017;9(1):e069.
92. Pedersen AB, Mikkelsen EM, Cronin-Fenton D, Kristensen NR, Pham TM, Pedersen L, et al. Missing data and multiple imputation in clinical epidemiological research. *Clin Epidemiol.* 2017;9:157–66.
93. Marston L, Carpenter JR, Walters KR, Morris RW, Nazareth I, Petersen I. Issues in multiple imputation of missing data for large general practice clinical databases. *Pharmacoepidemiol Drug Saf.* 2010;19:618–26.

94. Marston L, Carpenter JR, Walters KR, Morris RW, Nazareth I, White IR, et al. Smoker, ex-smoker or non-smoker? The validity of routinely recorded smoking status in UK primary care: A cross-sectional study. *BMJ Open*. 2014;4:e004958.
95. Sibley LM, Moineddin R, Agha MM, Glazier RH. Risk adjustment using administrative data-based and survey-derived methods for explaining physician utilization. *Med Care*. 2010;48(2):175–82.
96. van Buuren S, Groothuis-Oudshoorn K. MICE: multivariate imputation by chained equations in R. *J Stat Softw*. 2011;45(3):1–67.
97. Statistics Canada. Table 2, Canadian Tobacco, Alcohol and Drugs Survey (CTADS) [Internet]. 2017 [cited 2019 May 1]. Available from: <https://www.canada.ca/en/health-canada/services/canadian-tobacco-alcohol-drugs-survey/2017-summary/2017-detailed-tables.html#t2>
98. Gagné T. Estimation of smoking prevalence in Canada: Implications of survey characteristics in the CCHS and CTUMS/CTADS. *Can J Public Health*. 2017;108(3):e331–4.
99. Bushnik T, Hennessy DA, McAlister FA, Manuel DG. Factors associated with hypertension control among older Canadians. *Health Rep*. 2018;29(6):3–10.
100. Schindler-Ruwisch JM, Abroms LC, Bernstein SL, Heminger CL. A content analysis of electronic health record (EHR) functionality to support tobacco treatment. *Transl Behav Med*. 2017;7(2):148–56.

101. Taggar JS, Coleman T, Lewis S, Szatkowski L. The impact of the Quality and Outcomes Framework (QOF) on the recording of smoking targets in primary care medical records: Cross-sectional analyses from The Health Improvement Network (THIN) database. *BMC Public Health*. 2012;12:329.
102. IOM (Institute of Medicine). A nationwide framework for surveillance of cardiovascular and chronic lung diseases. Washington DC: The National Academies of Press; 2011.
103. Centre for Surveillance Coordination. Chronic disease surveillance in Canada: A background paper. Ottawa, Canada; 2003.
104. Lix L, Yogendran M, Shaw S, Burchill C, Metge C, Bond R. Population-based data sources for chronic disease surveillance. *Chronic Dis Can*. 2008;29(1):31–8.
105. German RR, Lee LM, Horan JM, Milstein RL, Pertowski CA, Waller MN. Updated guidelines for evaluating public health surveillance systems. Recommendations from the Guidelines Working Group. *Morb Mortal Wkly Rep*. 2001;50(RR13):1–35.
106. Goff DC, Brass L, Braun LT, Croft JB, Flesch JD, Fowkes FGR, et al. Essential features of a surveillance system to support the prevention and management of heart disease and stroke. *Circulation*. 2007;115:127–55.
107. Birkhead GS, Klompas M, Shah NR. Uses of electronic health records for public health surveillance to advance public health. *Annu Rev Public Health*. 2015;36:345–59.
108. Campbell NRC, McAlister FA, Quan H. Monitoring and evaluating efforts to control hypertension in Canada: Why, how, and what it tells us needs to be done about current care gaps. *Can J Cardiol*. 2013;29(5):564–70.

109. Canadian Institute for Health Information. Supply, distribution and migration of physicians in Canada, 2018 - data tables [Internet]. 2018 [cited 2019 Dec 10]. Available from: <https://secure.cihi.ca/estore/productSeries.htm?pc=PCC34>
110. Quan H, Khan N, Hemmelgarn BR, Tu K, Chen G, Campbell N, et al. Validation of a case definition to define hypertension using administrative data. *Hypertension*. 2009;54(6):1423–8.
111. Schwartz KL, Wilton AS, Langford BJ, Brown KA, Daneman N, Garber G, et al. Comparing prescribing and dispensing databases to study antibiotic use: A validation study of the Electronic Medical Record Administrative data Linked Database (EMRALD). *J Antimicrob Chemother*. 2019;74(7):2091–7.
112. Tu K, Wang M, Jaakkimainen RL, Butt D, Ivers NM, Young J, et al. Assessing the validity of using administrative data to identify patients with epilepsy. *Epilepsia*. 2014;55(2):335–43.
113. Tu K, Wang M, Young J, Green D, Ivers NM, Butt D, et al. Validity of administrative data for identifying patients who have had a stroke or transient ischemic attack using EMRALD as a reference standard. *Can J Cardiol*. 2013;29(11):1388–94.
114. Coleman N, Halas G, Peeler W, Casclang N, Williamson T, Katz A. From patient care to research: a validation study examining the factors contributing to data quality in a primary care electronic medical record database. *BMC Fam Pract*. 2015;16:11.
115. Auditor General of Alberta. Better healthcare for Albertans: a report by the Office of the Auditor General of Alberta [Internet]. Edmonton, AB; 2017. Available from: https://www.oag.ab.ca/documents/41/2017_-_Better_Healthcare_for_Albertans_Report_-_May_2017.pdf

116. Alberta Health Services. Connect Care [Internet]. 2020 [cited 2020 Apr 1]. Available from: <https://www.albertahealthservices.ca/info/cis.aspx>
117. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. *Health Inf Sci Syst*. 2014;2(1):3.
118. Benchimol EI, Smeeth L, Guttman A, Harron K, Moher D, Petersen I, et al. The reporting of studies conducted using observational routinely-collected health data (RECORD) statement. *PLOS Med*. 2015;12(10):e1001885.
119. Stewart M, Ryan B. Ecology of health care in Canada. *Can Fam Physician*. 2015;61:449–53.

**APPENDIX A: SUPPLEMENTARY TABLE 1: COMPARISON OF PATIENT
CHARACTERISTICS IN THE PROVINCIAL HEALTHCARE REGISTRY
AND THE CPCSSN PRIMARY CARE EMR DATABASE FOR ALBERTA**

Supplementary Table 1. Comparison of patient characteristics in the provincial healthcare registry and the CPCSSN primary care EMR database for Alberta.

	Administrative Data (general population)	Primary care EMR Data (all patients in database)
Data source	2017/18 AHCIP Registry ^a	CPCSSN data for Alberta up to June 30, 2018; patients with at least 1 visit in last 2 years
Total adults, N	3,481,340*	204,825
Female, n (%)	1,731,372 (49.7)	115,351 (56.3)
Male, n (%)	1,749,968 (50.3)	89,428 (43.7)
Age, mean (SD)	<i>not available</i>	49.7 (18.3)
Age groups, n (%)		
20-39 years	1,401,383 (40.3)	63,153 (30.8)
40-59 years	1,235,444 (35.5)	70,545 (34.4)
60-69 years	462,792 (13.3)	33,786 (16.5)
70-79 years	240,733 (6.9)	19,755 (9.6)
80 years and older	140,988 (4.0)	12,055 (5.9)
Urban, n (%)	4,184,418** (91.0)	159,737 (78.0)
Rural, n (%)	413,186** (9.0)	39,389 (19.2)
Missing/unknown residence or postal code, n (%)	485 (0.01)	5699 (2.8)

*Adults 20 years and older

**Urban/rural calculated based on second digit of postal code of each Local Geographic Area for discrete patients (all ages); denominator: 4,598,089

^a Alberta Health. Alberta Health Care Insurance Plan Statistical Supplement 2017/2018 [Internet]. 2018. Available from: https://open.alberta.ca/dataset/3c9a0637-29c1-4cb2-93ba-c2ac090ab2b5/resource/bca9cd6e-803d-439c-be69-b15a8dd800a4/download/ahcip-statistical-supplement-2017_2018_final.pdf

AHCIP=Alberta Health Care Insurance Plan; CPCSSN=Canadian Primary Care Sentinel Surveillance Network; EMR=electronic medical record; SD=standard deviation

APPENDIX B: COPYRIGHT PERMISSIONS

Creative Commons Legal Code

Attribution 4.0 International

Official translations of this license are available [in other languages](#).

Creative Commons Corporation (“Creative Commons”) is not a law firm and does not provide legal services or legal advice. Distribution of Creative Commons public licenses does not create a lawyer-client or other relationship. Creative Commons makes its licenses and related information available on an “as-is” basis. Creative Commons gives no warranties regarding its licenses, any material licensed under their terms and conditions, or any related information. Creative Commons disclaims all liability for damages resulting from their use to the fullest extent possible.

Using Creative Commons Public Licenses

Creative Commons public licenses provide a standard set of terms and conditions that creators and other rights holders may use to share original works of authorship and other material subject to copyright and certain other rights specified in the public license below. The following considerations are for informational purposes only, are not exhaustive, and do not form part of our licenses.

Considerations for licensors: Our public licenses are intended for use by those authorized to give the public permission to use material in ways otherwise restricted by copyright and certain other rights. Our licenses are irrevocable. Licensors should read and understand the terms and conditions of the license they choose before applying it. Licensors should also secure all rights necessary before applying our licenses so that the public can reuse the material as expected. Licensors should clearly mark any material not subject to the license. This includes other CC-licensed material, or material used under an exception or limitation to copyright.

Considerations for the public: By using one of our public licenses, a licensor grants the public permission to use the licensed material under specified terms and conditions. If the licensor’s permission is not necessary for any reason—for example, because of any applicable exception or limitation to copyright—then that use is not regulated by the license. Our licenses grant only permissions under copyright and certain other rights that a licensor has authority to grant. Use of the licensed material may still be restricted for other reasons, including because others have copyright or

other rights in the material. A licensor may make special requests, such as asking that all changes be marked or described. Although not required by our licenses, you are encouraged to respect those requests where reasonable.

Creative Commons Attribution 4.0 International Public License

By exercising the Licensed Rights (defined below), You accept and agree to be bound by the terms and conditions of this Creative Commons Attribution 4.0 International Public License ("Public License"). To the extent this Public License may be interpreted as a contract, You are granted the Licensed Rights in consideration of Your acceptance of these terms and conditions, and the Licensor grants You such rights in consideration of benefits the Licensor receives from making the Licensed Material available under these terms and conditions.

Section 1 – Definitions.

- a. **Adapted Material** means material subject to Copyright and Similar Rights that is derived from or based upon the Licensed Material and in which the Licensed Material is translated, altered, arranged, transformed, or otherwise modified in a manner requiring permission under the Copyright and Similar Rights held by the Licensor. For purposes of this Public License, where the Licensed Material is a musical work, performance, or sound recording, Adapted Material is always produced where the Licensed Material is synched in timed relation with a moving image.
- b. **Adapter's License** means the license You apply to Your Copyright and Similar Rights in Your contributions to Adapted Material in accordance with the terms and conditions of this Public License.
- c. **Copyright and Similar Rights** means copyright and/or similar rights closely related to copyright including, without limitation, performance, broadcast, sound recording, and Sui Generis Database Rights, without regard to how the rights are labeled or categorized. For purposes of this Public License, the rights specified in Section [2\(b\)\(1\)-\(2\)](#) are not Copyright and Similar Rights.
- d. **Effective Technological Measures** means those measures that, in the absence of proper authority, may not be circumvented under laws fulfilling obligations under Article 11 of the WIPO Copyright Treaty adopted on December 20, 1996, and/or similar international agreements.
- e. **Exceptions and Limitations** means fair use, fair dealing, and/or any other exception or limitation to Copyright and Similar Rights that applies to Your use of the Licensed Material.
- f. **Licensed Material** means the artistic or literary work, database, or other material to which the Licensor applied this Public License.
- g. **Licensed Rights** means the rights granted to You subject to the terms and conditions of this Public License, which are limited to all Copyright and Similar Rights that apply to Your use of the Licensed Material and that the Licensor has authority to license.

- h. **Licensors** means the individual(s) or entity(ies) granting rights under this Public License.
- i. **Share** means to provide material to the public by any means or process that requires permission under the Licensed Rights, such as reproduction, public display, public performance, distribution, dissemination, communication, or importation, and to make material available to the public including in ways that members of the public may access the material from a place and at a time individually chosen by them.
- j. **Sui Generis Database Rights** means rights other than copyright resulting from Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases, as amended and/or succeeded, as well as other essentially equivalent rights anywhere in the world.
- k. **You** means the individual or entity exercising the Licensed Rights under this Public License. **Your** has a corresponding meaning.

Section 2 – Scope.

a. License grant.

1. Subject to the terms and conditions of this Public License, the Licensor hereby grants You a worldwide, royalty-free, non-sublicensable, non-exclusive, irrevocable license to exercise the Licensed Rights in the Licensed Material to:
 - A. reproduce and Share the Licensed Material, in whole or in part; and
 - B. produce, reproduce, and Share Adapted Material.
2. Exceptions and Limitations. For the avoidance of doubt, where Exceptions and Limitations apply to Your use, this Public License does not apply, and You do not need to comply with its terms and conditions.
3. Term. The term of this Public License is specified in Section 6(a).
4. Media and formats; technical modifications allowed. The Licensor authorizes You to exercise the Licensed Rights in all media and formats whether now known or hereafter created, and to make technical modifications necessary to do so. The Licensor waives and/or agrees not to assert any right or authority to forbid You from making technical modifications necessary to exercise the Licensed Rights, including technical modifications necessary to circumvent Effective Technological Measures. For purposes of this Public License, simply making modifications authorized by this Section 2(a)(4) never produces Adapted Material.
5. Downstream recipients.
 - A. Offer from the Licensor – Licensed Material. Every recipient of the Licensed Material automatically receives an offer from the Licensor to exercise the Licensed Rights under the terms and conditions of this Public License.
 - B. No downstream restrictions. You may not offer or impose any additional or different terms or conditions on, or apply any Effective Technological Measures to, the Licensed Material if

doing so restricts exercise of the Licensed Rights by any recipient of the Licensed Material.

6. No endorsement. Nothing in this Public License constitutes or may be construed as permission to assert or imply that You are, or that Your use of the Licensed Material is, connected with, or sponsored, endorsed, or granted official status by, the Licensor or others designated to receive attribution as provided in Section 3(a)(1)(A)(i).

b. Other rights.

1. Moral rights, such as the right of integrity, are not licensed under this Public License, nor are publicity, privacy, and/or other similar personality rights; however, to the extent possible, the Licensor waives and/or agrees not to assert any such rights held by the Licensor to the limited extent necessary to allow You to exercise the Licensed Rights, but not otherwise.
2. Patent and trademark rights are not licensed under this Public License.
3. To the extent possible, the Licensor waives any right to collect royalties from You for the exercise of the Licensed Rights, whether directly or through a collecting society under any voluntary or waivable statutory or compulsory licensing scheme. In all other cases the Licensor expressly reserves any right to collect such royalties.

Section 3 – License Conditions.

Your exercise of the Licensed Rights is expressly made subject to the following conditions.

a. Attribution.

1. If You Share the Licensed Material (including in modified form), You must:
 - A. retain the following if it is supplied by the Licensor with the Licensed Material:
 - i. identification of the creator(s) of the Licensed Material and any others designated to receive attribution, in any reasonable manner requested by the Licensor (including by pseudonym if designated);
 - ii. a copyright notice;
 - iii. a notice that refers to this Public License;
 - iv. a notice that refers to the disclaimer of warranties;
 - v. a URI or hyperlink to the Licensed Material to the extent reasonably practicable;
 - B. indicate if You modified the Licensed Material and retain an indication of any previous modifications; and
 - C. indicate the Licensed Material is licensed under this Public License, and include the text of,

or the URI or hyperlink to, this Public License.

2. You may satisfy the conditions in Section 3(a)(1) in any reasonable manner based on the medium, means, and context in which You Share the Licensed Material. For example, it may be reasonable to satisfy the conditions by providing a URI or hyperlink to a resource that includes the required information.
3. If requested by the Licensor, You must remove any of the information required by Section 3(a)(1)(A) to the extent reasonably practicable.
4. If You Share Adapted Material You produce, the Adapter's License You apply must not prevent recipients of the Adapted Material from complying with this Public License.

Section 4 – Sui Generis Database Rights.

Where the Licensed Rights include Sui Generis Database Rights that apply to Your use of the Licensed Material:

- a. for the avoidance of doubt, Section 2(a)(1) grants You the right to extract, reuse, reproduce, and Share all or a substantial portion of the contents of the database;
- b. if You include all or a substantial portion of the database contents in a database in which You have Sui Generis Database Rights, then the database in which You have Sui Generis Database Rights (but not its individual contents) is Adapted Material; and
- c. You must comply with the conditions in Section 3(a) if You Share all or a substantial portion of the contents of the database.

For the avoidance of doubt, this Section 4 supplements and does not replace Your obligations under this Public License where the Licensed Rights include other Copyright and Similar Rights.

Section 5 – Disclaimer of Warranties and Limitation of Liability.

- a. **Unless otherwise separately undertaken by the Licensor, to the extent possible, the Licensor offers the Licensed Material as-is and as-available, and makes no representations or warranties of any kind concerning the Licensed Material, whether express, implied, statutory, or other. This includes, without limitation, warranties of title, merchantability, fitness for a particular purpose, non-infringement, absence of latent or other defects, accuracy, or the presence or absence of errors, whether or not known or discoverable. Where disclaimers of warranties are not allowed in full or in part, this disclaimer may not apply to You.**
- b. **To the extent possible, in no event will the Licensor be liable to You on any legal theory (including, without limitation, negligence) or otherwise for any direct, special, indirect,**

incidental, consequential, punitive, exemplary, or other losses, costs, expenses, or damages arising out of this Public License or use of the Licensed Material, even if the Licensor has been advised of the possibility of such losses, costs, expenses, or damages. Where a limitation of liability is not allowed in full or in part, this limitation may not apply to You.

- c. The disclaimer of warranties and limitation of liability provided above shall be interpreted in a manner that, to the extent possible, most closely approximates an absolute disclaimer and waiver of all liability.

Section 6 – Term and Termination.

- a. This Public License applies for the term of the Copyright and Similar Rights licensed here. However, if You fail to comply with this Public License, then Your rights under this Public License terminate automatically.
- b. Where Your right to use the Licensed Material has terminated under Section 6(a), it reinstates:
 - 1. automatically as of the date the violation is cured, provided it is cured within 30 days of Your discovery of the violation; or
 - 2. upon express reinstatement by the Licensor.For the avoidance of doubt, this Section 6(b) does not affect any right the Licensor may have to seek remedies for Your violations of this Public License.
- c. For the avoidance of doubt, the Licensor may also offer the Licensed Material under separate terms or conditions or stop distributing the Licensed Material at any time; however, doing so will not terminate this Public License.
- d. Sections 1, 5, 6, 7, and 8 survive termination of this Public License.

Section 7 – Other Terms and Conditions.

- a. The Licensor shall not be bound by any additional or different terms or conditions communicated by You unless expressly agreed.
- b. Any arrangements, understandings, or agreements regarding the Licensed Material not stated herein are separate from and independent of the terms and conditions of this Public License.

Section 8 – Interpretation.

- a. For the avoidance of doubt, this Public License does not, and shall not be interpreted to, reduce,

limit, restrict, or impose conditions on any use of the Licensed Material that could lawfully be made without permission under this Public License.

- b. To the extent possible, if any provision of this Public License is deemed unenforceable, it shall be automatically reformed to the minimum extent necessary to make it enforceable. If the provision cannot be reformed, it shall be severed from this Public License without affecting the enforceability of the remaining terms and conditions.
- c. No term or condition of this Public License will be waived and no failure to comply consented to unless expressly agreed to by the Licensor.
- d. Nothing in this Public License constitutes or may be interpreted as a limitation upon, or waiver of, any privileges and immunities that apply to the Licensor or You, including from the legal processes of any jurisdiction or authority.

Creative Commons is not a party to its public licenses. Notwithstanding, Creative Commons may elect to apply one of its public licenses to material it publishes and in those instances will be considered the “Licensor.” The text of the Creative Commons public licenses is dedicated to the public domain under the [CC0 Public Domain Dedication](#). Except for the limited purpose of indicating that material is shared under a Creative Commons public license or as otherwise permitted by the Creative Commons policies published at creativecommons.org/policies, Creative Commons does not authorize the use of the trademark “Creative Commons” or any other trademark or logo of Creative Commons without its prior written consent including, without limitation, in connection with any unauthorized modifications to any of its public licenses or any other arrangements, understandings, or agreements concerning use of licensed material. For the avoidance of doubt, this paragraph does not form part of the public licenses.

Creative Commons may be contacted at creativecommons.org.

Additional languages available: العربية, čeština, Deutsch, Ελληνικά, Español, euskara, suomeksi, français, hrvatski, Bahasa Indonesia, italiano, 日本語, 한국어, Lietuvių, latviski, te reo Māori, Nederlands, norsk, polski, português, русский, svenska, Türkçe, українська, 中文, 華語. Please read the [FAQ](#) for more information about official translations.

Subject: Re: Your permission needed for manuscript
Date: Monday, May 11, 2020 at 2:21:45 PM Mountain Daylight Time
From: Hude Quan
To: Stephanie Garies

Yes permitted

Hude

From: Stephanie Garies
Sent: May 10, 2020 1:15 PM
To: Hude Quan
Subject: FW: Your permission needed for manuscript

Hi all,

I will be finalizing my thesis shortly, which includes the following manuscripts listed below; these have not yet been published and therefore I need permission from each co-author to publish these in my thesis document. Could you reply to this email with your permission to do so?

1. Garies S, McBrien K, Quan H, Manca D, Drummond N, Williamson T. A data quality assessment to inform hypertension surveillance using primary care electronic medical record data. Submitted November 2019 to *BMC Public Health*
2. Garies S, Youngson E, Soos B, Forst B, Duerksen K, Manca D, McBrien K, Drummond N, Quan H, Williamson T. Primary care EMR and administrative data linkage in Alberta, Canada: describing the suitability for hypertension surveillance. Submitted April 2020 to *BMJ Health & Care Informatics*

Thanks,
Stephanie

Subject: Re: Your permission needed for manuscript

Date: Wednesday, May 6, 2020 at 1:05:58 PM Mountain Daylight Time

From: Donna Manca

To: Stephanie Garies

Hi Stephanie,

Of course I approve!

Best Regards,
Donna

On May 6, 2020, at 10:32 AM, Stephanie Garies

wrote:

Hi all,

I will be finalizing my thesis shortly, which includes the following manuscripts listed below; these have not yet been published and therefore I need permission from each co-author to publish these in my thesis document. Could you reply to this email with your permission to do so?

1. Garies S, McBrien K, Quan H, Manca D, Drummond N, Williamson T. A data quality assessment to inform hypertension surveillance using primary care electronic medical record data. Submitted November 2019 to
2. Garies S, Youngson E, Soos B, Forst B, Duerksen K, Manca D, McBrien K, Drummond N, Quan H, Williamson T. Primary care EMR and administrative data linkage in Alberta, Canada: describing the suitability for hypertension surveillance. Submitted April 2020 to

Thanks,
Stephanie

Subject: Re: Your permission needed for manuscript
Date: Wednesday, May 6, 2020 at 12:42:14 PM Mountain Daylight Time
From: Tyler Williamson
To: Stephanie Garies, Kerry Alison McBrien
CC: Neil Drummond, Hude Quan, Donna Manca

Hi Stephanie,

I happily approve this!

Tyler

On May 6, 2020, 11:14 AM -0600, Kerry Alison McBrien

wrote:

Hi Stephanie,

I have no problem with this.

Thanks,
Kerry

On May 6, 2020, at 10:32 AM, Stephanie Garies

wrote:

Hi all,

I will be finalizing my thesis shortly, which includes the following manuscripts listed below; these have not yet been published and therefore I need permission from each co-author to publish these in my thesis document. Could you reply to this email with your permission to do so?

1. Garies S, McBrien K, Quan H, Manca D, Drummond N, Williamson T. A data quality assessment to inform hypertension surveillance using primary care electronic medical record data. Submitted November 2019 to *BMC Public Health*
2. Garies S, Youngson E, Soos B, Forst B, Duerksen K, Manca D, McBrien K, Drummond N, Quan H, Williamson T. Primary care EMR and administrative data linkage in Alberta, Canada: describing the suitability for hypertension surveillance. Submitted April 2020 to *BMJ Health & Care Informatics*

Subject: Re: Your permission needed for manuscript

Date: Wednesday, May 6, 2020 at 10:40:48 AM Mountain Daylight Time

From: Neil Drummond

To: Stephanie Garies

I do so permit.

Neil Drummond

On Wed, 6 May 2020 at 10:32, Stephanie Garies

wrote:

Hi all,

I will be finalizing my thesis shortly, which includes the following manuscripts listed below; these have not yet been published and therefore I need permission from each co-author to publish these in my thesis document. Could you reply to this email with your permission to do so?

1. Garies S, McBrien K, Quan H, Manca D, Drummond N, Williamson T. A data quality assessment to inform hypertension surveillance using primary care electronic medical record data. Submitted November 2019 to *BMC Public Health*
2. Garies S, Youngson E, Soos B, Forst B, Duerksen K, Manca D, McBrien K, Drummond N, Quan H, Williamson T. Primary care EMR and administrative data linkage in Alberta, Canada: describing the suitability for hypertension surveillance. Submitted April 2020 to *BMJ Health & Care Informatics*

Thanks,

Stephanie

--

Neil Drummond PhD
Professor and Alberta Health Services Chair in Primary Care Research
University of Alberta
Department of Family Medicine

Adjunct Professor, School of Public Health, University of Alberta;

Adjunct Professor, Departments of Family Medicine and Community Health Sciences, University of Calgary

Subject: RE: Your permission needed for manuscript

Date: Wednesday, May 6, 2020 at 10:40:49 AM Mountain Daylight Time

From: Erik Youngson

To: Stephanie Garies

I have no objection to this.

Thanks,

Erik

From: Stephanie Garies

Sent: Wednesday, May 6, 2020 10:33 AM

Subject: Your permission needed for manuscript

Caution - This email came from an external address and may contain unsafe content. Ensure you trust this sender before opening attachments or clicking any links in this message.

Hi all,

I will be finalizing my thesis shortly, which will include the following manuscript listed below; this has not yet been published and therefore I need permission from each co-author to publish this in my thesis document. Could you reply to this email with your permission to do so?

1. Garies S, Youngson E, Soos B, Forst B, Duerksen K, Manca D, McBrien K, Drummond N, Quan H, Williamson T. Primary care EMR and administrative data linkage in Alberta, Canada: describing the suitability for hypertension surveillance. Submitted April 2020 to *BMJ Health & Care Informatics*

Thanks,
Stephanie

This message and any attached documents are only for the use of the intended recipient(s), are confidential and may contain privileged information. Any unauthorized review, use, retransmission, or other disclosure is strictly prohibited. If you have received this message in error, please notify the sender immediately, and then delete the original message. Thank you.

Subject: Re: Your permission needed for manuscript

Date: Wednesday, May 6, 2020 at 11:45:43 AM Mountain Daylight Time

From: Brian Forst

To: Stephanie Garies

Yes, you have my permission to publish the paper in

Brian

On Wed, May 6, 2020 at 10:32 AM Stephanie Garies ·

wrote:

Hi all,

I will be finalizing my thesis shortly, which will include the following manuscript listed below; this has not yet been published and therefore I need permission from each co-author to publish this in my thesis document. Could you reply to this email with your permission to do so?

1. Garies S, Youngson E, Soos B, Forst B, Duerksen K, Manca D, McBrien K, Drummond N, Quan H, Williamson T. Primary care EMR and administrative data linkage in Alberta, Canada: describing the suitability for hypertension surveillance. Submitted April 2020 to *BMJ Health & Care Informatics*

Thanks,

Stephanie

Subject: Re: Your permission needed for manuscript

Date: Wednesday, May 6, 2020 at 10:41:12 AM Mountain Daylight Time

From: Kimberley Duerksen

To: Stephanie Garies

Yes, I approve. Good luck!

On Wed, May 6, 2020 at 10:32 AM Stephanie Garies

wrote:

Hi all,

I will be finalizing my thesis shortly, which will include the following manuscript listed below; this has not yet been published and therefore I need permission from each co-author to publish this in my thesis document. Could you reply to this email with your permission to do so?

1. Garies S, Youngson E, Soos B, Forst B, Duerksen K, Manca D, McBrien K, Drummond N, Quan H, Williamson T. Primary care EMR and administrative data linkage in Alberta, Canada: describing the suitability for hypertension surveillance. Submitted April 2020 to *BMJ Health & Care Informatics*

Thanks,

Stephanie

--

Kimberley Duerksen, MSc., IAPP.
Research Coordinator
Department of Family Medicine
University of Alberta

Subject: Re: Your permission needed for manuscript
Date: Wednesday, May 6, 2020 at 10:42:03 AM Mountain Daylight Time
From: Boglarka Soos
To: Stephanie Garies

Hi Stephanie,

Congratulations on working toward finalizing your thesis.
You have my permission to cite the publication.

Larka

--

Boglarka Soos
Data Manager
Canadian Primary Care Sentinel Surveillance Network (CPCSSN)
Strategy for Patient Oriented Research (SPOR) Primary and Integrated Health Care Innovation
Network (PIHCIN)
Southern Alberta Primary Care Research Network (SAPCRen)

Department of Family Medicine
University of Calgary

From: Stephanie Garies
Date: Wednesday, May 6, 2020 at 10:32 AM

Subject: Your permission needed for manuscript

Hi all,

I will be finalizing my thesis shortly, which will include the following manuscript listed below; this has not yet been published and therefore I need permission from each co-author to publish this in my thesis document. Could you reply to this email with your permission to do so?

1. Garies S, Youngson E, Soos B, Forst B, Duerksen K, Manca D, McBrien K, Drummond N, Quan H, Williamson T. Primary care EMR and administrative data linkage in Alberta, Canada: describing the suitability for hypertension surveillance. Submitted April 2020 to *BMJ Health & Care Informatics*

Thanks,
Stephanie