6-9-2022 2:30 PM

Development of a Multi-Factorial Data Quality Score for Primary Care Electronic Medical Records

Kathryn Stirling, *The University of Western Ontario*

Supervisor: Ryan, Bridget, *The University of Western Ontario*
Joint Supervisor: Terry, Amanda, *The University of Western Ontario*
A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Epidemiology and Biostatistics

Abstract

As healthcare providers transitioned from paper-based records to electronic medical records (EMRs), researchers gained access to more patient and care data, with greater detail. Current research regarding Canadian primary care EMR data suggests the quality of data is variable. For researchers who wish to use EMR data, it is important to have a method of evaluating data quality that is applicable to multiple EMR datasets so research quality can be assured. There is currently no unified scoring system for primary care EMR data. This thesis built on previous EMR data quality research by developing and testing a composite measure of data quality using previously-validated quality measures that assessed the data quality domains of completeness, correctness, and currency. A composite data quality score was created and tested using data splitting. This scoring system could be used by researchers to examine EMR data quality and compare data quality across data sources.

Keywords: Electronic Medical Record, Electronic Health Record, Data Quality, Primary Care

Summary for Lay Audience

As healthcare providers transitioned from paper-based medical record systems to electronic medical records (EMRs), researchers gained access to more patient and care data, with a greater level of detail because the data were available in electronic formats, which means there was no longer the need for manual reviews of hundreds of paper charts. Current research regarding Canadian primary care EMR data suggests that the quality of data is variable. This variation can be caused by EMRs having different input requirements, different ways of storing data, and differences in how individuals use the EMR. For researchers who wish to use EMR data, it is important to have a method of evaluating data quality that is applicable to multiple EMR datasets. The quality of conclusions drawn by research studies depends on many factors including the quality of the data used. However, researchers often do not assess the quality of the EMR data they use. There is currently no unified scoring system for primary care EMR data; a common scoring system would make the assessment of primary care EMR data easier. This thesis built on previous EMR data quality research by developing and testing a composite measure of data quality. Previously developed and tested measures of data quality were used. These measures assessed the data quality domains of completeness, correctness, and currency. A composite data quality score was created by generating values of data quality for different aspects within each of the data quality domains and then combining these values using addition and averaging. The score's reliability was tested by splitting the data into two parts and then using the first part to create the score and the second part to replicate the process. The score was found to be reliable across the two groups. This scoring system could be used by researchers to examine EMR data quality and compare data quality across data sources.

Acknowledgements

I would like to first and foremost thank my supervisors Dr. Bridget Ryan and Dr. Amanda Terry. Without their help and guidance none of this work would have been possible. I am extremely grateful for their ongoing support throughout this thesis and for their compassionate mentorship.

Thank you as well to my thesis committee members Dr. Tyler Williamson and Dr. Alex Singer and my thesis defence committee members Dr. Mark Speechley, Dr. Shehzad Ali, and Dr. Richard Booth. The perspectives they brought to this thesis from their areas of expertise encouraged me to dig deeper into the material and take a thoughtful approach to the questions of data quality.

I would also like to thank Jason Black, Rick Truant, Vijaya Chevendra, and Sylvia Aponte-Hao for their assistance with data management and advice and guidance throughout the data processing steps of this project. Their skill and knowledge in the areas of data management and manipulation, and their willingness to answer questions were vital to this thesis; without them I would have been lost in the sea of EMR data.

Finally I would like to thank my family and friends who have supported and cheered me on through the thesis process and who have indulged me by listening to my monologues about data quality, EMRs, and the pitfalls of coding. Their care and encouragement have and continue to be invaluable.

Table of Contents

List of Tables

List of Figures

List of Appendices

Chapter 1

1 Introduction

This thesis focuses on assessing the quality of electronic medical record (EMR) primary care data for use in research. EMRs are digital health records, containing a portion of a person's health information over their life, which are held by health care providers[1]. Different EMR systems are in use within Canada and differ even within single provinces. Each EMR system has a particular way to enter, store and extract data resulting in considerable variation across different systems. Additionally, the way individual health care providers use their EMR systems introduces further variation[2]. While the focus of this thesis is on data quality for research use, assessing EMR data quality is essential for many reasons. Within the primary care context, health care providers rely on EMR data to support quality patient care[3]. Some EMRs include decision support functions whose functionality depends on quality data input[4]. Further, the data in EMRs can be used to monitor the quality of patient care and outcomes accurately[5]. In a policy and research context, data quality will directly affect the quality of the final product whether that is scientific discovery or public policy. Previous research into Canadian primary care EMR data suggests that the quality of data is variable[6–8]. For researchers who are interested in using EMR data, it is essential to have a method of evaluating data quality that is applicable to multiple EMR datasets, so that conclusions drawn from the research conducted are reliable.

Data quality in EMRs has previously been conceptualized as spanning four domains: comparability, completeness, correctness, and currency/timeliness[9]. Other researchers have defined three domains of EMR data quality which are essential: comparability, completeness and correctness[10]. In the context of Canadian primary care research, seven domains of EMR data quality have been identified[11], which have significant overlap with the previously discussed domains. The present thesis draws most from a study by Terry et al. focused specifically on Canadian primary care EMR data[8]. Terry et al. assessed data quality using the domains of comparability, completeness, correctness, and currency/timeliness individually for three different EMR datasets[8].

Previous EMR data quality research has focused on single domains[6,12] and even when multiple domains are examined, the scores are evaluated separately[8]. For ease of use by researchers and to allow interpretability of the overall quality of the data, a single composite score would be useful. The current thesis sought to create a composite score of primary care EMR data quality that can be calculated prior to using the data for research purposes.

1.1 Thesis Objectives

The primary objective of this thesis was to create a composite data quality score that would be easily implemented by researchers using primary care EMR data. This was achieved by turning the domains of data quality into executable steps, calculating scores on assessment methods, combining these assessment methods into domain scores, and combining domain scores into a composite data quality score. The same methods were used on a second subset of the data to determine replicability of the scoring method. A secondary objective of this thesis, achieved as part of the primary objective, was to create scores for each of the domains of data quality examined in this thesis (completeness, correctness, and currency) so that comparisons could be made on domain data quality as well as overall data quality.

This thesis begins with Chapter 1, an introduction to the thesis. Chapter 2 reviews the literature concerning EMR data quality and data quality assessment, particularly as it applies to primary care research. Chapter 3 describes the methods employed in meeting the objectives of the thesis. Chapter 4 describes the results. Chapter 5 discusses the strengths and limitations of this thesis, places the results of this thesis within the context of the body of literature and discusses the utility of this research for informing primary care research that uses EMR data, particularly as it relates to the importance of researchers examining and reporting on the quality of those data.

Chapter 2

2 Literature Review

This chapter provides an overview of the current research into electronic medical record data quality, both the actual data quality and how that data quality is conceptualized, with a focus on electronic medical records used in primary care practice.

2.1 Electronic Medical Records (EMRs)

There are several terms which may be used to refer to digital health records. The most common terms in the Canadian context are electronic medical records (EMRs), electronic health records (EHRs), and personal health record (PHRs). Though there are many similarities, and these terms are occasionally used interchangeably in the literature, disambiguation is important for an accurate understanding.

EMRs are partial health records, containing a portion of a person's health information over their life, which are held by health care providers[1]. In Canada, the term EMR is also used to refer to the software product used by a physician to maintain patient health records[1]. EHRs are complete health records, containing all of a person's health information over their life, which are held by health care providers[1]. However, in the current literature the terms EMR and EHR are sometimes used interchangeably, with the common definition which is here attributed to EMR. PHRs are complete or partial health records that are not held by health care providers (i.e. they are in the custody of the patient or a family member) [1]. Though the findings of the present thesis may be more widely applicable to other types of data, the data source used in this study was derived from EMRs and therefore the main focus here is on EMRs used in Canadian primary care settings.

2.2 EMR Use in Primary Care

EMRs have become very common in Canadian primary care. Based on a 2019 survey of Canadian primary care physicians, 86% use an EMR, though this percentage varies significantly across the provinces and territories. This number increased from 73% in 2015[13]. Younger physicians and those who worked in a group practice were more likely to use an EMR[13]. However, the more complex functionalities of EMR systems may be underutilized; as an example, less than half of surveyed primary care physicians in Canada routinely used a computer system, like an EMR, to support quality of care decisions (reminders for follow-up care, guideline based screenings, or sending patients with test results)[13]. This observed pattern of underuse could be related to the quality or useability of the tools themselves, a lack of time on the part of physicians to learn how to use such tools, or even poor functionality of the tools themselves which rely on the assumption of high-quality data to function properly.

Rubinowicz and colleagues surveyed clinicians in 2016 in primary care clinics across Canada where EMRs had been implemented three to eight years previously[14]. Of these surveyed clinicians, who had an EMR for several years, 87% used electronic prescribing routinely and almost 78% used electronic prompts regarding potential dosing or interaction hazards[14]. Family physicians surveyed reported that EMRs allowed better monitoring of patient progress, more efficient billing, and improved quality and continuity of care[14]. Most clinicians surveyed by Rubinowicz et al. also felt that EMR implementation had reduced waiting times, care costs, and risk of errors[14].

Another survey of 331 family physicians in 2015 in Québec who used an EMR regularly in their practice showed slightly different results[15]. Physicians surveyed used on average 67% of their EMR software's clinical functions (e.g. clinical notes, electronic prescribing), 47% of the communication functions (e.g. tracking test results, referrals), and 90% of the administrative functions (e.g. billing, scheduling). It was posited by Raymond and colleagues that some of the underuse of EMR functions may be due to physicians not knowing that such features exist or not understanding how to use them[15]. Alternately, this underuse could be related to how the EMR software is designed and how quick and easy such features are to use[15].

Given the increase in EMR usage since 2015, it is possible that the differences seen across these studies are related to the changes in adoption and comfort with EMRs over the past few years. Increased use of an EMR, better ease of use, and greater physician satisfaction with an EMR were found to correlate with performance benefits including better efficiency and quality of care[15]. Ease of EMR use was also positively correlated with physician satisfaction with the EMR[15].

2.3 How EMR Data Quality Has Been Explored Previously

Clinicians using EMRs are doing so in order to provide patient care, and so, while EMR-derived data have been used for research, to inform policy decisions, and to make decisions regarding changes to patient care standards, these data were not created for these purposes[5,16–18]. Therefore, the usefulness of EMR data for any application depends on the quality of the data itself, the ability to extract the data from the EMR, and the structure of the data once it is extracted[19]. Fitness for purpose is an aspect of data quality which has been conceptualized as "the property of data produced by a measurement process that enables a user of the data to make technically correct decisions for a stated purpose"[20]. In the research environment, frequently this is stated as the ability of data to answer a given research question.

Data quality in EMRs has been conceptualized in several different ways. When examining data quality there are four commonly used concepts which apply at different levels: 1) frameworks; 2) domains (sometimes referred to as dimensions); 3) measures (sometimes referred to as metrics); and 4) data quality assessment methods. Frameworks are the overarching models of data quality within which are contained domains[9]. Frameworks differ in the number of domains they include, and which domains are used. Frameworks may be specific to a purpose or more generally oriented to data quality as a whole. Domains of data quality are broad conceptual classifications of the aspects of data, which can contribute to quality.

Weiskopf and Weng identified five potentially relevant domains of data quality: concordance, completeness, correctness, plausibility and currency/timeliness[9]. Measures, also referred to as metrics, are more specific criteria of data quality, which are usually sub-categories of domains. For example within the domain of currency, the measure of "timeliness of antenatal care" has been used[8]. Data quality assessment methods are ways that the measures are operationalized [9]. For example, the previous measure of "timeliness of antenatal care" used the assessment method "Percentage of patients with a positive pregnancy laboratory test result and one or more visits within two months of the result"[8].

The domains of data quality described by Weiskopf and Weng are similar to those described by the United Nations Statistical Commission's National Quality Assurance Framework, which is a framework meant to aid countries in the development of national quality assurance frameworks that will standardize national statistics and ensure quality and trustworthiness of these data[21]. The National Quality Assurance Framework defines data quality in terms of five general domains: relevance, accuracy and reliability, timeliness and punctuality, coherence and comparability, accessibility and clarity[21]. Relevance is the degree to which the information meets the needs (both present and potential/emerging) of the users. Relevance includes dimensions of completeness, user needs, and user satisfaction[21]. Accuracy and reliability indicate how closely the information reflects the true values it is measuring and how consistently the information reflects reality over time. Accuracy is often broken down into sampling and non-sampling error and then further into systematic and random error of which systematic error is of greater concern due to its potential to distort the data in ways which make the results of analyses inaccurate[21]. Timeliness refers to how quickly the data are made available after collection and punctuality refers to whether the data were made available on the prespecified date[21]. Accessibility and clarity indicate the extent of information, including metadata (the data which provides information about the actual data being used) and supplemental explanatory information (e.g. source of the data, collection and processing methods), which is readily accessible[21]. Coherence and comparability indicate the internal consistency of the data and the comparability over time and geography, the use of common or standard metrics and facilitation of combination of data from disparate sources[21].

Kahn et al. created a framework intended to serve as a common terminology for data quality researchers[22]. This framework was focused on data quality intrinsic-to-data elements, for example the data's format or distributions, rather than extrinsic-to-data elements, such as data accessibility or determining fitness for use. The framework organized terms from the data quality literature into two data quality assessment contexts, verification and validation, and three data quality categories, conformance, completeness, and plausibility[22]. The data quality assessment contexts differ in that verification relates to measuring data quality based on intrinsic factors, no external data or reference points are used, whereas validation measures the data quality extrinsically and therefore uses external comparators to assess the data[22]. The data quality assessments developed in this thesis fall under the verification assessment context as no external comparators are needed to generate the scores.

A scoping review of data quality research, with a specific focus on identifying and codifying commonly used data quality dimensions and assessment methods, was recently conducted by Bian et al. The review found 14 data quality dimensions and 10 data quality assessment methods, though some overlap of data quality dimensions was described[23]. The most commonly focused on data quality dimensions in the frameworks and studies analysed were: completeness, concordance, correctness/ accuracy, and plausibility[23]. Completeness was defined as the presence, or absence, of data regardless of data values[22]. Concordance was defined as the agreement of elements within the data or the agreement of elements between the data and an external source[9]. Correctness/ accuracy was defined as the truth of the values present in the data[9]. Plausibility was defined as the credibility of the values in the data within the context of other present variables, measured for example by comparing the actual and expected relationship of two variables[22]. The most commonly used assessment methods in the reviewed literature were: element presence, data source agreement, validity check, and data element agreement. Element presence was defined as whether the expected elements are present in the data[9]. Data source agreement was defined as the level to which the data in the dataset being examined matches data from an external source[9,22]. Validity check was defined as whether values in the dataset agree with data from an external source or with self-evident concepts, and the degree to which time-varying elements vary as expected or as seen in external sources[22]. Data element agreement was defined as the degree to which data elements within the dataset of interest contain compatible values[9,22]. It was posited in the review that some data quality dimensions and assessment methods are more commonly used because they are easier to understand or easier to operationalize, though this does not mean the less frequently used dimensions and assessment methods are less important to understanding in evaluating data quality than the more commonly used ones[23]. The review stresses that, though there have been attempts, there is not yet a widespread adoption of data quality framework use[23]. Further, the importance of reporting on data quality is described as essential to ensuring transparency and consistency in research and to enabling comparison of data quality across the wide variety of datasets available to researchers[23].

A review that looked at data quality assessment of real-world data more generally, rather than a specific focus on EMR data, was conducted by Liaw et al[24]. Although this review examined a more diverse study base (compared to the literature review by Bian et al. which focused on clinical data) there are still concepts Liaw et al. addresses which are relevant to EMR data. The review divided data quality concepts into intrinsic, contextual and technical factors[24]. Intrinsic factors included concepts such as completeness and correctness[24]. Contextual factors included concepts like the reputation of the data source and the relevance of the data to the research question[24]. Technical factors included concepts related to data linkage and processing[24]. In terms of this thesis the most relevant are the intrinsic data quality factors. Liaw et al.'s review stresses the necessity of examining and reporting on data quality at all points across the data life cycle including the importance of researchers routinely reporting data quality[24].

A recent framework developed to model data quality for Canadian primary care EMR data used four domains of data quality: comparability, completeness, correctness, and currency/timeliness[8]. Comparability refers to how consistent EMR data are with an external population[9]. Completeness refers to how much of what is observed (during a patient visit for example) is actually recorded within the EMR[25]. Correctness refers to the accuracy of the data recorded in the EMR[25]. Currency refers to whether the data in the EMR are up to date for a given time point[9]. Other researchers have also used the same concepts, comparability, completeness and correctness, to define data quality[10]. In the context of Canadian primary care research, some researchers have used seven domains of EMR data quality[11] which have significant overlap with the domains in the previously discussed frameworks.

While many different combinations of data quality domains have been used in previous EMR data quality assessment research, this thesis used three of the five core constructs of data quality identified by Weiskopf and colleagues[9]: completeness, correctness, and currency/timeliness. These domains map well onto key components of the United Nations Statistical Commission's National Quality Assurance Framework[21] and are also suitable for EMR data. While previous EMR data quality investigations have examined comparability, this domain will not be included in the present analysis. Comparability is omitted because measures of comparability are highly population-specific and therefore are not applicable to a generalizable scoring system.

2.4 EMR Data Quality Considerations

Large easily accessible datasets, such as those generated by the increasing digitization of all kinds of patient health data through EMR usage, create opportunities for researchers, policy makers and clinicians[16]. For example, researchers have the opportunity to develop and answer research questions with sample sizes that would be unfeasible with primary data collection due to financial or time constraints. The use of data analytics in health care has the potential to increase quality of care, lower costs, and lead to better outcomes by leveraging this newly accessible data to answer existing questions which may have been too expensive or obscure previously[16].

In clinical research, these data sets allow for large scale adverse event monitoring, conducting studies that can lead to improved quality of care, and the creation of statistical algorithms that can directly impact both research design and execution[16]. For public health, large digital datasets allow fast monitoring of disease patterns and analysis of large-scale trends that can lead to targeted assessments and interventions[16]. However, there are many challenges and potential pitfalls to be aware of when using these datasets. The massive volume of data available to researchers, and the speed at which more data becomes available is novel in terms of the conduct of modern research[16]. Though this speed can be leveraged; for example, to facilitate real-time tracking of infectious disease outbreaks, it requires new and innovative research designs. These types of data also present the challenge of differences in format and structure. Data collected from different electronic sources, even just different EMRs, will have differences in formatting. Parts of the data may be structured or semi-structured; however, others will be unstructured, such as in clinical notes, creating challenges in compilation and standardization[16].

Perhaps the most important challenge to the use of large health datasets is the quality of the data themselves. Health research has direct impacts on patient care, meaning it has the potential to change patients' lives positively or negatively. If research is based on poor quality data, erroneous conclusions can be reached which can have far-reaching impacts on policy and care decisions. Unfortunately, health care data, especially unstructured health data, often has errors[16]. In massive data sets, errors, especially systematic ones, can be amplified; hence, reinforcing why researchers must always be mindful that EMR-derived data were not created for research purposes.

As the usage of EMRs expands further in Canadian primary care, opportunities are created for researchers to examine various questions and patterns of primary care in local, provincial, and national contexts[19]. Primary care EMRs can be used for chronic disease surveillance. This may lead to better prediction modelling and risk assessment for these diseases[19]. The widespread availability of EMR data means that not only can disease patterns be examined but the overall health system can also be evaluated[19]. By compiling multiple EMR data sets, provincial and local health systems can be evaluated and improved. Currently, much of this type of assessment is based on health care administrative data[19]. Health care administrative data include data on drugs, physician services (e.g. diagnoses and procedures), hospital services (e.g. diagnoses, information about hospital stays, procedures), and patient identification (e.g. IDs, birth dates, postal codes)[26]. However, EMR data contain clinical information previously only available through chart review, an expensive and time-consuming process. These EMR data provide more detail than health care administrative data in many areas, for example patient information and disease progression over time, and offer the opportunity to assess different data quality dimensions including quality of care[19].

2.5 Data Quality Deficits in EMR Data

Despite the importance of high data quality in EMRs, previous studies have indicated quality deficits in EMR-derived data. A cohort study using EMRs from Dutch primary care practices was conducted to determine the data quality for breast, colon, and prostate cancer diagnoses. The completeness, correctness, and concordance with the reference standard (in this case the national Netherlands Cancer registry) were assessed. Data quality was quantified as standard incidence ratios, the ratio of observed cases in the EMR database to expected cases based in the reference standard multiplied by 100%[12]. While the overall standard incidence ratios were relatively good, there was large variability observed across cancer types, across time points, and across different EMRs[12]. Overall, up to 30% of cases for the cancers assessed could be missed if the EMR data alone were used[12]. Further, a large number of false positives were identified in the data[12]. One of the largest sources of variability identified was which type of EMR was used[12]. This variability indicates that researchers should include EMR type as a variable in their statistical models when using data consolidated from multiple EMRs as it may significantly impact findings.

In another study, the completeness of problem lists, the area of a medical record where a patient's relevant medical history is entered, was assessed in relation to chronic diseases in EMRs used in 119 primary care practices in Manitoba[27]. Completeness was assessed by comparing the number of diagnoses documented on EMR problems lists with the number of billings related to each disease[27]. The average completeness varied by disease but also by primary care provider[27]. While the completeness percentage for some conditions was relatively high; for example, 80% for diabetes, the low end of completeness was 43% for COPD, which is indicative of a serious deficit[27]. This lack of completeness in a key area of the EMR data demonstrates the potential for EMRs to produce low quality data that, if used indiscriminately, could lead to erroneous research conclusions.

An Australian study examined the data quality of EMR data from primary care clinics to assess whether these data would be appropriate to use in a prediction model for knee replacement surgery in patients with osteoarthritis[28]. The variables to be used in the model were assessed for completeness and plausibility. Data were also compared to external data sources to assess accuracy and validity[28]. Completeness and accuracy of some key variables, year of knee replacement, side of knee replacement, and year of death, were found to be very low[28]. However, completeness and plausibility for the predictor variables chosen, including prescriptions and other diagnoses, were high with the exception of low completeness values for BMI and weight gain over time[28]. The mixed data quality between variables and domains indicates the need to assess the quality of EMR data prior to use.

EMRs provide clinicians with the opportunity to assess patient care across their practice, potentially identifying areas for improvement[19]. In one example of this, EMR data was leveraged to incorporate a screening tool (the Screening Tool of Older People's Prescriptions) into EMRs to prevent potentially inappropriate prescriptions[29]. After the implementation of the screening tool, there were no changes in potentially inappropriate prescribing in the implementation group[29]. One of the possible explanations proposed for this was that the EMRs had data quality issues, specifically around data completeness and mis- or underuse of coded fields[29]. When physicians were interviewed, they were unaware of the deficits in data quality in their EMRs[29]. Physicians also did not consider that not using or misusing data entry portions of the EMR could have downstream effects, including impacting the utility of built-in decision support tools[29]. This lack of consideration of the downstream impact of EMR data or lack of data is perhaps unsurprising given that the focus of clinicians is on the patient and their care which may not align with collecting and recording the maximum amount of data. This mis-/ disuse of the EMR fields may also indicate a design flaw of the EMR.

Reliability, defined as data being recorded consistently across time and recorders, is an aspect of data quality that is essential to ensuring accuracy. Data from 18 primary care physicians in Toronto Ontario who used EMRs were assessed for reliability[7]. Reliability was measured by first calculating the proportion of patients eligible for a Pap smear, mammogram, or influenza vaccine, who were recorded in the EMR as receiving this service over a given time period[7]. The change in these calculated rates was then compared to the change in rates from administrative data, used as the reference standard, regarding the same procedures over the same time period[7]. Differences were found in the changes measured by the EMR data compared to the administrative datasets, which is indicative of unreliability in the EMR data[7].

Current research into Canadian primary care EMR data suggests that the quality of this data is variable[6–8]. Previous research on Canadian primary care EMR data quality assessed comparability, completeness, correctness, and currency/timeliness individually for three different EMR datasets[8]. The data used were obtained from the DELPHI database which is a research database of de-identified EMR data from 18 primary care practice sites in south western Ontario and is based at the Centre for Studies in Family Medicine at Western University in London, Ontario[30].

Comparability of the DELPHI patient population to the Canadian census population was found to be high in one dataset but significant differences between the census population and the EMR populations were found in the other two datasets[8]. Completeness varied significantly across the three datasets and by measure and test condition. Test conditions were specific health conditions of interest chosen based on clinical expertise, their prior use in EMR research, and their frequent occurrence in primary care practice[8]. Correctness also varied significantly by test condition and across datasets with positive predictive values of just 4% for obesity in one data set to 80% for diabetes in another dataset[8]. However, the presence of unlikely combinations of age-specific procedures was found to be 0% in all datasets[8]. Currency also varied across the datasets[8]. Although not all measures were appropriate for all three datasets, the results demonstrate EMR datasets vary in quality and therefore data quality assessment is essential.

2.6 Creation of Composite Scores

In addition to assessing the individual components of data quality, as has been done in previous studies[6–8,10,25], there is utility in creating composite scores from the individual components. Composite scores would allow for comparability across projects. Having a single composite score would be easier for researchers to use and therefore could encourage the examination of data quality in projects where data quality is not the main focus.

There are a number of methods described in the literature for combining multiple scores from component metrics into a single composite score that is both useful and interpretable. Combining variables into a single composite must be done carefully as it can: change the strength of relationships with variables not used in the creation of the component score (e.g. outcomes); obscure important information; and make interpretation more difficult[31]. Methods of weighting that can be used when creating composite scores include weights based on expert opinion, empirical regression weights, weights based on previous studies/data collected, unit/standardized weights, geometric means, arithmetic means, summation, and radar charts. However, there is no standard practice in the clinical literature for combining measurements on different scales that are used to measure different aspects of the same concept[32]. Therefore, several methods were explored to determine their suitability to the particular case of combining individual data quality metrics to obtain a single score that can accurately describe the quality of an EMR-derived dataset, as was done in this thesis.

In expert weighting, experts in the field are consulted as to which scoring components are the most important or relevant to the research question being asked in a specific project. This method allows tailoring of the score to the particular needs of a project and therefore might be preferred; for example, if a project's question is highly dependent on the currency/timeliness of the data, the timeliness score might be weighted more highly to provide a score that reflects the project priorities. Expert opinion was used by Terry et al. in initial component score development[8]. However, it is not appropriate to apply expert opinion to the creation of weights for the combined score in this thesis. In the case of individual studies wanting a data quality score tailored to their specific objectives, expert opinion could be used; however, this thesis sought to create a more universal composite score that would be comparable across various research questions. Weights derived from expert opinion would necessarily differ based on the research question being explored and therefore would not be comparable.

Regression weighting – deriving weights based on least squares regression analysis – is only an option if it is possible to use a single dependent variable, and this variable is known, and if the sample size is sufficiently large, and sampling error is low[33]. Regression weights are calculated by maximizing the linear relationship between the predictors and the dependent variable[33]. In the case of developing a single composite score of EMR data quality, it is not possible to use regression weighting because the dependent variable (i.e. the final combined score) does not yet exist.

The weights used to create the final composite EMR data quality score also cannot be derived from previous studies or other data as there are no studies which have previously combined the individual data quality scores in the way that is proposed here. Further, there are no previous studies which have used the score calculation method developed in Terry et al[8].

The term 'unit weights' is used by Bobko et al. to refer to standardizing scores; for example, by converting each individual component variable to Z-scores and then applying equal weighting[33]. 'Unit weights' has also been used in the literature to refer to the summation of raw or unstandardized scores, but this definition will not be used here. Song and colleagues define the same concept, summing the Z-scores of the original variables that will make up the composite score, as simple averaging[31]. Unit weights allow adjustment for differing variances of the raw component scores which can otherwise significantly impact a composite score[33]. However, in the present thesis, the format of the individual component scores is not compatible with a Z-scores transformation. This is because the individual component scores are calculated by examining the percentage of patients in certain groups in the EMR; for example, one assessment method for completeness is the percentage of patients in the EMR who have at least one entry in their allergy record (including “no allergies”). There is no mean or standard deviation which can be calculated for this value when examining a single EMR database and therefore Z-score based methods are inappropriate.

Unit weighting relies on the scores either being normally distributed or being transformed to be normally distributed and for all the scores to have the same directional relationship with the underlying concept[31–33].

Gerstein and colleagues suggest the simpler method of calculating the geometric mean of the measurements to be combined as an alternative to unit weighting. Using a geometric mean also requires all the component measurements to have near normal distributions (or to be transformed to normal), to have the same directional relationship to the underlying construct, and to be positive and greater than zero[32]. The geometric mean is calculated by taking the *n*th root of the product of the *n* components[32]. The geometric mean can be used to combine measurements from different scales with different distributions because it reflects the relationships between the components rather than their absolute values[32]. In the case of this thesis, it is not a useful option because it is possible to have a score of zero on an assessment method. In the development of the assessment methods by Terry and colleagues scores of zero were calculated for sensitivity and positive predictive value for asthma and for unlikely combinations of age and specific procedure[o].

An alternative to the geometric mean is the simple arithmetic mean, derived by adding *n* components and then dividing the total by *n*. The arithmetic mean is a commonly used method of combining values and is suitable to use when combining percentages, where giving equal weighting to component measurements is not a concern. If some components have a higher number of component measurements than others, this is not accounted for in the arithmetic mean. Unlike the geometric mean, values of zero are not of concern when calculating the arithmetic mean.

Summation of values is a step in the process of calculating an arithmetic mean but can also be used as a method independent of the mean calculation to combine values into a composite score. Summation weights each value in the sum; in essence each value is given equal weight in the final score[34]. This type of weighting implies that each value within the score should contribute equally to the final score, which is also true of mean scores[34]. Summation provides simple to explain and interpret scores. However, summation does not allow comparison of scores which have different possible maximums due to having different numbers of values contributing to the totals.

Radar charts are a method of data visualization which is particularly useful when displaying multi-dimensional data where all variables use the same scale[35]. In the case of EMR data quality, all the measures are presented as percentages which makes radar charts a convenient option to display all of the examined aspects of quality in a single and easily understandable way. The calculation of the area of a filled radar chart has sometimes been used to create a composite score. However, this is not an appropriate method of composite score creation because the ordering of the axis on the chart can influence the area significantly and therefore artificially impact the composite measure[36]. For example, if there is a dataset with two axes with very high scores and two axes with very low scores in a dataset placing the high scores opposite each other will generally create a smaller area that placing the high scoring axes beside each other.

2.7 Considerations Specific to Developing Composite Data Quality Scores for Primary Care EMR Research

In order to examine whether the measures of EMR data quality that were developed and employed in this thesis were relevant to the type of research being done using primary care EMR data, a review was conducted of research that has been conducted using the Canadian Primary Care Sentinel Surveillance Network (CPCSSN). CPCSSN is the data source for the present thesis and contains EMR data which is moderately representative of the Canadian population with patients being slightly older and more likely to be female than the Canadian population, and physicians included in CPCSSN being younger and more likely to be female than the total population of primary care providers in Canada[37]. CPCSSN compiles a frequently updated list of published literature utilizing CPCSSN-held EMR data. This list was the basis of this review. As of June 1st, 2021, this list contained 143 articles. Of these articles, 102 used the CPCSSN data and indicated in their publication which data fields or data tables were used.

The majority of studies reviewed used patient demographics (sex: 89 studies and age: 93 studies) in their research. The other most used components of the CPCSSN data were the problems list (the health conditions with which a patient has been diagnosed): 80 studies, billing data: 72 studies, CPCSSN Disease Case table (disease cases identified by the CPCSSN diagnostic algorithms based on the contents of patient records): 67 studies, the encounter diagnosis table: 66 studies, and the medications table: 65 studies. A summary of all reported data use in the CPCSSN studies can be found in Appendix A.

Based on this review, some components of the CPCSSN data were found to be used more in primary care EMR-based research than other components were. In order to get the highest utility from a composite data quality assessment score, it makes sense to focus on metrics that use the most used components of the data and conversely to either exclude or place less weight on components of the data which are less widely used. One example of a lesser used component of the data are vaccination records where only one of 102 studies used patient vaccination records. This indicates that relying heavily on data from the vaccination table to create a widely useable score might not be the best fit. An alternate explanation for the disparities in use could be that the tables which are used less frequently are the tables known to have lower quality data and therefore researchers are more hesitant to include them in their research.

2.8 Caveats in the Evaluation of Data Quality

When evaluating EMR data quality it is important to acknowledge that there are, at present, no gold standards for how the established domains of data quality are operationalized. Data quality refers to both the inherent quality of the data but also its fitness for use in each particular situation where one considers employing data. This multi-faceted nature and associated data quality metrics, therefore, cannot function as black and white indicators which deem the data “good” or “bad”. Depending on factors such as the desired end use for the data, the origin of the data, and the variables and information contained within the dataset, the domains of data quality can be operationalized in multiple ways. The methods of operationalizing data quality domains, including measures and assessment methods that are presented in this thesis, were derived from the work of Terry and colleagues[8]. These data quality measures were developed through an iterative process that included review of the EMR data quality literature, examination of structured EMR data, assessment of the measures in combination with each other, and creation of assessment methods[8].

Similarly, in the literature from which this thesis drew the data quality assessment methods, the process of developing and operationalizing the data quality measures was extensive and included consultation of expert opinion and evaluation of multiple possible combinations of potential data quality measures[8]. It is important to note that these measures are only one of many potential ways to operationalize and measure EMR data quality. While they are not necessarily the ideal depiction of the quality of the data, they represent the results of an exacting process that balanced practicality and rigorousness, consistent with research in this area[8,24]. When developing data quality measurement tools, choices and trade-offs must be made in the attempt to capture the multi-faceted nature of data quality while still having a useable data quality tool. The work of Terry and colleagues, and by extension this thesis, attempted to develop a helpful and useable method of measuring data quality of data derived from primary care EMRs.

2.9 Contribution of This Thesis to EMR Data Quality Assessment

While primary care data collected from EMRs are a significant asset to researchers, data re-use for research purposes can cause harm if not done judiciously[12]. Though they are frequently used in research already, primary care EMR data were not designed for this purpose and therefore the research using these data should be carefully designed, keeping the data source and quality in mind[12].

This thesis builds on previous EMR data quality research by developing and replicating a single composite score of data quality using three previously validated quality domains that assessed completeness, correctness, and currency/timeliness individually[8]. This scoring system once established, could be used by research teams to assess the quality of data for research, where quality issues are identified, to be potentially corrected. The scoring system created in this thesis has the benefit of being a simple and straightforward method of identifying potential data quality issues in a user-friendly manner. Currently, few researchers using EMR derived data in their research report anything about the quality of the data that were used[23]. It is hoped that the simplicity of the composite scores will encourage more researchers to critically examine their data, report on the quality of data use in a transparent way, and may prompt further and more in-depth investigation of potential issues and strengths of EMR-derived data.

While the primary purpose of the composite score developed in this thesis is to assist researchers in one aspect of their examination of the data used in their research, this score has other potential applications. A single score that can be straightforwardly calculated will allow data aggregators, such as CPCSSN, to look at trends in data quality over time and across dimensions such as the EMR provider and the practice-based research networks that contribute data to CPCSSN. Because factors such as data cleaning algorithms (procedures used to standardize and streamline the data) and input prompts (prompts given to the clinician as they enter data into the EMR) change over time and differ between networks, it would be expected that a composite data score would show differences based on these factors. By introducing a simple score for examining data quality, the impacts of such changes to the data could be quantified. Further, a consistent scoring system could assist EMR designers in creating products that are optimized not only as a user interface for care providers but that also optimize the quality of data which can be collected from the EMRs for research.

Chapter 3

3 Methods

3.1 Data Source: The Canadian Primary Care Sentinel Surveillance System (CPCSSN)

The source of data for this research was the Canadian Primary Care Sentinel Surveillance Network (CPCSSN)[38]. The dataset was obtained through CPCSSN by submitting a data request and project description after ethics approval had been obtained for the thesis from the Western University Health Sciences Research Ethics Board (Ethics #115903).

CPCSSN is a pan-Canadian primary care research initiative which aims to improve quality of care for Canadians through securely collecting and reporting on data from electronic medical records (EMRs)[38]. The CPCSSN database contains over 1.8 million de-identified patient-level records. The data come from physicians participating in 10 practice-based research networks across Canada, extracted from multiple EMR software packages[38]. CPCSSN uses algorithms to clean and validate the collected EMR data[39]. These validation methods and data cleaning algorithms have been published in peer-reviewed journals and are available through the CPCSSN website [38,40]. This cleaning results in data that are much easier to use in research; however, data can still have quality issues stemming from their original input at point of care. The CPCSSN database includes clinical information (encounter diagnoses, prescribed medications, and procedures), practice site characteristics (e.g. EMR type, site location), provider characteristics (e.g. provider type, sex), and patient characteristics (e.g. sex, potential risk factors)[38]. The data cleaning can result in new data quality issues in some instances as data cleaning is not identical across CPCSSN networks. These differences in data cleaning mean that different networks may have different data quality. This adds an additional layer of data quality consideration as data quality can be impacted at the patient/provider level, at the EMR level, at the CPCSSN network level, and then at the CPCSSN project level when all of the network data is merged. Each of these points can differently impact data quality and each merge of data is a new opportunity for issues to be introduced to the data.

Table 1 describes the CPCSSN data structure. Patient ID is a unique random number assigned by CPCSSN to a patient in order to identify that patient across the database and link all other variables associated with that patient; that is, externally useless. Encounter ID is the unique randomly generated numeric code assigned to a particular patient encounter recorded in the database which allows linking of all events related to a single encounter (e.g. a patient might receive a physical examination, diagnosis, and medication prescription in a single encounter). Patients often have multiple entries in a single variable category. For example, a patient may have more than one health condition, which may have been entered on multiple dates. These multiple entries can be linked through the associated patient ID and encounter ID.

The data used for this thesis were stored on a Microsoft SQL Server in the Western DELPHI EMR Database project secure server environment, located inside Western University's firewalls. A Microsoft SQL Server stores information in tables; each table is comprised of rows and columns. The overall structure of a SQL database is a series of these linked tables. Two Microsoft SQL Server concepts, important to this thesis' analyses were primary keys and foreign keys. When a column is denoted as a primary key, it indicates the data in that column is unique (meaning all of the values in the column are different, there is no repetition) and there are no null values; for example, in the CPCSSN dataset the 'PatientID' column in the 'Patient' table is a primary key. When a column is denoted as a foreign key, it indicates the column is a link between tables. The foreign key column is referencing a column from another table, frequently a primary key; for example, in the CPCSSN dataset, the 'PatientID' column in any table other than the ‘Patient’ table is a foreign key. These concepts were important for the analyses because table linkage allowed patient records to be followed across tables. Several analyses, for example looking at age-specific procedures, required linking data in multiple tables to a single patient record.

Prior to the statistical analysis, the CPCSSN dataset was split into two randomly selected equally sized groups of patients, referred to as the score Development Group and the score Replication Group, so that the data quality score could be developed in one group and then replicated in a second group. Data splitting is a common practice in algorithm development and can assist in examining reliability[41,42] Each group contained all variables. Cases were assigned to each group by randomly assigning each unique patient ID which is attached to every record in the dataset either a 0 (score Development Group) or a 1 (score Replication Group) which ensured: no patient in the dataset was in both groups or was without a group; an even group split; and easy separation of groups during analysis.

Table 1: CPCSSN tables and columns used in the composite score creation

CPCSSN Database Tables			
Patient	Health Condition	Exam	Encounter
Patient ID Column: - Unique patient identification number Birthyear Column: - Patient birthyear Sex Column: - Patient Sex (male or female) Postal Code: - Rural/ urban designation of the patient postal code	Patient ID Column: - Unique patient identification number Encounter ID Column: - Unique encounter identification number Diagnosis Text Column: - The text of the diagnosis as it was entered in the EMR Diagnosis Code Type Column: - Type of code (e.g. ‘ICD9’) Diagnosis Code Column: - The diagnosis code corresponding to the diagnosis (e.g. ‘250.0’) Date of Onset Column: - The date the health condition began	Patient ID Column: - Unique patient identification number Encounter ID Column: - Unique encounter identification number Exam 1 Column: - Name of exam (e.g. ‘BMI’, ‘dBP’, ‘sBP’) Result 1 Column: - Result of the exam Exam 2 Column: - Name of paired exam, can only be dBP (Exam 1 must be sBP) Result 2 Column: - Result of the exam	Patient ID Column: - Unique patient identification number Encounter ID Column: - Unique encounter identification number Encounter Date Column: - The date an encounter occurred

CPCSSN Tables Continued				
Medication	Lab	Vaccination	Allergy	Billing
Patient ID Column: - Unique patient identification number Encounter ID Column: - Unique encounter identification number Start Date Column - Date the patient started taking the medication Stop Date Column - Date the patient stopped taking the medication Name Column: - Name of the medication	Patient ID Column: - Unique patient identification number Encounter ID Column: - Unique encounter identification number Performed Date Column: - Date the lab test was done Name Column: - Name of the lab test Test Result Column: - Result of the lab test	Patient ID Column: - Unique patient identification number Encounter ID Column: - Unique encounter identification number Given Date Column: - Date the vaccination was administered Name Column: - Name of the vaccine	Patient ID Column: - Unique patient identification number Encounter ID Column: - Unique encounter identification number Name Column: - Name of the allergy (can also be “no allergies”)	Patient ID Column: - Unique patient identification number Encounter ID Column: - Unique encounter identification number Service Date Column: - Date the billing was performed/submitted Diagnosis Code Type: - Type of code (e.g. ‘ICD9’) Diagnosis Code Column: - The diagnosis code corresponding to the diagnosis (e.g. ‘250.0’)

3.2 Outcomes

While the main objective of this thesis was to create a single composite score representative of data quality, there were four outcomes in total created for this thesis, with the purpose of providing metrics that can be used to assess data quality in primary care EMR-derived datasets. Consistent with the domains used by Terry et al.[8], three domain-level scores were created to measure data quality concerning completeness, correctness, and currency/timeliness. The fourth domain, comparability, requires comparison to an outside dataset and is highly population-specific and therefore is not applicable to a generalizable scoring system. One composite score was created that combined the three domain-level metrics.

3.3 Data Examination and Preparation

When using complex data such as found in the CPCSSN dataset, it is necessary to examine the data closely to ensure the statistical methods used are appropriate. To this end, data were examined to determine what variables existed in the data set and in what form these variables were presented as well as how the tables within the Microsoft SQL server (on which the data were stored) related to each other. This process involved reviewing the CPCSSN-provided data dictionary, which gave the table and column names, the data type for each column, and brief descriptions of the contents of each column. The CPCSSN entity relationship diagram was also reviewed. The entity relationship diagram provides an overview of how each of the tables within the SQL database are connected to each other, and of the primary and foreign keys in the database that are used to create the connections. A version of the CPCSSN entity relationship diagram is available through the CPCSSN website[43].

The variables used in the analyses are described in

, including the format the variables were given in the database.

Table 2: Variables used in analysis

Variable Name	Description	Response Format
Patient Birthyear	Year a patient was born	Year
Patient Sex	Sex of a patient	"Male" or "Female"
Patient's geographic area	Indication whether a patient lives in an urban or rural area	"U' if urban, "R" if rural
Diagnosis Code	The diagnosis code which was entered in the EMR problem list for a diagnosis given to a particular patient. In the billing table it refers to the code submitted to the province as billing data for a patient encounter.	ICD9 code
Encounter Date	The date a given patient-provider encounter occurred	Year-Month-Day
Exam (Present or Absent)		
Systolic Blood Pressure (sBP)		Present or Absent
Diastolic Blood Pressure (dBP)		Present or Absent
Height		Present or Absent
Weight		Present or Absent
Body Mass Index (BMI)		Present or Absent
Exam		
Systolic Blood Pressure (sBP)		mmHg
Diastolic Blood Pressure (dBP)		mmHg
Height		cm
Weight		kg
Body Mass Index (BMI)		kg/m^2
Medication Prescribed	Name of the medication as recorded by the prescribing physician	Free text
Medication Start Date	Date the patient started taking the medication	Year-Month-Day
Medication Stop Date	Date the patient stopped taking the medication	Year-Month-Day
Lab	There are several subtypes of lab test	Present or absent

Variable Name	Description	Response Format
Beta-hCG (pregnancy test)	The level of Beta-human chorionic gonadotropin in the blood, a marker of pregnancy	mIU/mL
Fasting plasma glucose	Blood glucose levels in a patient who has been fasting. Can be used as a method of diagnosing diabetes mellitus.	mmol/L
Random plasma glucose	Blood glucose levels in a patient who has not been fasting. Can be used as a method of diagnosing diabetes mellitus.	mmol/L
Oral glucose tolerance test	Blood glucose in a patient 2 hours after consuming a 75-g glucose solution orally. Also referred to as a glucose challenge test. Can be used as a method of diagnosing diabetes mellitus.	mmol/L
HbA1c level	Measures the percentage of glycated haemoglobin in the blood. Can be used as a method of diagnosing or monitoring diabetes mellitus.	A1c value (%)
Thyroid stimulating hormone (TSH)	Measures the amount of TSH in the blood. Can be used to diagnose hypothyroidism.	uIU/L
Lab performed date	The date a given lab test was performed	Year-Month-Day
Allergy Name	The name of an allergy	Free text
Vaccination Name	The name of a vaccination that was given to a patient	Free text
Vaccination Date	The date a specific vaccination was given to a patient.	Year-Month-Day

3.4 Preliminary and Descriptive Analysis

Data were analysed using R statistical software (V4.0.3; R Core Team 2021) through R-studio. Patient demographics were examined for the overall dataset as well as the two data subsets (score Development Group and score Replication Group). Distributions of patient age (age distribution by decade), sex, and patient location (rural or urban), were reported. Comparisons between the Development and Replication Groups to assess comparability between the groups were made using t-tests to compare the birth year distributions and Chi square tests to compare sex and location distributions. Due to the very large sample size, small but statistically significant differences were expected between the Development and Replication groups;[44] therefore, effect size estimates (Cohen's d and Cramer's V respectively) were also calculated and reported.

3.5 Creation of Composite Scores

In the creation of a composite score, it is necessary to first consider theoretical factors. The purpose of the scoring methodology created in this thesis was to provide a single number that will reflect data quality. This score of data quality in this thesis is conceived as the combination of the completeness, correctness, and currency/timeliness of the data. These domains all reference different aspects of data quality and all refer to the multidimensional overarching construct of data quality as a whole. The basis for combining the individual scores into the composite is the assumption that the domains are representative of differing aspects of the construct of data quality[9,31]. It is reasonable to make the assumption that the domains are all referring to the overarching construct of data quality because the assessment methods and measures which compose them were developed specifically to measure individual aspects of the construct of data quality.

Two score creation methods were chosen to create domain-level scores and the composite score: summation and calculation of the arithmetic mean. In simple summation, each of the assessment method scores were summed together to form the domain scores which were then summed to a total composite score. This method was chosen as the simplest way of combining the scores. It is also an easily understandable method for anyone who might use the score and allows comparison of the domain scores and the overall score between different projects or datasets. Summation does not allow comparison of the domains within a single assessment however because the theoretical maximum summation scores are different for each domain. The arithmetic mean is calculated by dividing the sum of the n components and dividing that by n. The arithmetic mean has the advantage of allowing comparison of domain scores within a project or dataset; for example, comparing the completeness and correctness scores to each other within a project, as well as comparisons of domain and overall scores between different projects or datasets.

3.6 Hierarchy of Data Quality

As conceived in this thesis, the highest level of data quality assessment is the composite score. The composite score is made up of domain-specific scores where domains refer to the overarching components of data quality that have previously been defined in data quality research[8,9,11] with the domains of completeness, correctness, and currency used in this thesis. Measures are the individual components of data quality that are used to define a particular domain. Multiple assessment methods are used to capture a particular measure. A given assessment method calculates a numerical metric from the data. As an example, for the domain *Completeness*, one measure of data quality is *Consistency of capture*. One assessment method that assesses the *Consistency of capture* is the percentage of patients that have one or more problem list entries.

Table 3 provides this hierarchy of data quality assessment as executed in this thesis. All individual data quality measures and corresponding assessment methods used in this thesis were adapted from the paper by Terry and colleagues which used EMR-derived data[8] that were similar to the data used here. In the paper by Terry et al., 11 data quality measures were operationalized to cover the four domains of data quality being examined (comparability, completeness, correctness, and currency/timeliness)[8]. These operationalized data quality measures cover the commonly used components of data found in the review of previous CPCSSN studies conducted in this thesis. This thesis adapts nine of these measures because the present thesis is focused only on completeness, correctness, and currency/timeliness; the two measures previously used to examine comparability were not included.

Several of the data quality assessment methods were applicable, and therefore calculated, only for patients with relevant health conditions. These six health conditions are referred to as *test conditions* and were chosen by Terry and colleagues based on clinical expertise, prior use in EMR research, and frequent occurrence in primary care practice[8] and are: diabetes, hypertension, hypothyroidism, asthma, obesity, and urinary tract infection. The case definitions for these six test conditions can be found in Appendix B.

Table 3: Summary of components of the composite score

CPCSSN Table Used	Assessment Method	Measure	Domain	Composite Score
Health Condition, Medication, Lab, Billing	Sensitivity value for diabetes mellitus (test condition definition vs. billing code)	Sensitivity	Completeness	Composite Score
Health Condition, Encounter, Exam, Medication, Billing	Sensitivity value for hypertension (test condition definition vs. billing code)			
Health Condition, Medication, Lab, Billing	Sensitivity value for hypothyroidism (test condition definition vs. billing code)			
Health Condition, Medication, Billing	Sensitivity value for asthma (test condition definition vs. billing code)			
Health Condition, Exam, Billing	Sensitivity value for obesity (test condition definition vs. billing code)			
Health Condition, Medication, Billing	Sensitivity value for urinary tract infection (test condition definition vs. billing code)			
Patient, Health Condition	% of patients with 1 or more problem list entries	Consistency of capture		
Patient, Allergy	% of patients with 1 or more allergy record entries			
Encounter, Medication	% of patients who visited in past year with 1 or more prescribed meds			
Patient, Exam	% of patients with 1 or more blood pressure recordings for patients 18 + years	Recording of blood pressure, height, and weight		
Patient, Exam	% of patients with 1 or more height recordings			
Patient, Exam	% of patients with 1 or more weight recording			
Health Condition, Medication, Lab, Exam	% of patients with diabetes mellitus, with 1 or more blood pressure recordings	Recording of blood pressure among patients requiring a		

CPCSSN Table Used	Assessment Method	Measure	Domain	Composite Score
Health Condition, Encounter, Exam, Medication,	% of patients with hypertension medications (2 or more oral anti- hypertensives, or 1 or more diuretics) with 1 or more blood pressure recordings	blood pressure measurement		
Health Condition, Medication, Lab, Billing	Positive predictive value for diabetes mellitus (test condition definition vs. billing code)	Positive predictive value	Correctness	
Health Condition, Encounter, Exam, Medication, Billing	Positive predictive value for hypertension (test condition definition vs. billing code)			
Health Condition, Medication, Lab, Billing	Positive predictive value for hypothyroidism (test condition definition vs. billing code)			
Health Condition, Medication, Billing	Positive predictive value for asthma (test condition definition vs. billing code)			
Health Condition, Exam, Billing	Positive predictive value for obesity (test condition definition vs. billing code)			
Health Condition, Medication, Billing	Positive predictive value for urinary tract infection (test condition definition vs. billing code)			
Patient, Vaccination	% of patients ≥10 yrs. of age with a tetanus toxoid conjugate vaccine (diphtheria, haemophilus B, pertussis, polio, and tetanus)	Unlikely combinations of age & specific procedures		
Health Condition, Medication, Lab, Exam, Encounter	% of obese patients with 1 or more weight recordings within 1 year of last visit in recorded in the database	Timeliness of weight recordings for patients with obesity	Currency	
Lab, Encounter	% of patients with a positive pregnancy laboratory test result and 1 or more visits within two months of the result	Timeliness of visit for antenatal care		
Patient, Exam, Encounter	% of 18yr.+ patients with 1 or more blood pressure values recorded ≤ one year prior to their last visit in the database	Timeliness of blood pressure,		

CPCSSN Table Used	Assessment Method	Measure	Domain	Composite Score
Patient, Exam, Encounter	%of patients with 1 or more height values recorded ≤ one year prior to their last visit in the database	height, and weight recordings		
Patient, Exam, Encounter	% of patients with 1 or more weight values recorded ≤ one year prior to their last visit in the database for			

3.7 Domains and Their Associated Measures and Assessment Methods

Completeness Domain: Four measures were developed by Terry et al. to examine completeness: sensitivity, consistency of capture, recording of height and weight, and recording of blood pressure among patients requiring a measurement; i.e., those with diabetes mellitus and those taking hypertension medications[8]. The assessment method for Sensitivity required calculating sensitivity values for all six test conditions using the test condition definitions as a reference standard and billing codes as the comparison standard[8] (Figure 1, see Appendix D for example calculation) . For Consistency of capture, the assessment method was calculating the percentages of patients with one or more entries on their problems list, one or more entries in their allergy record, and patients visiting their care provider in the last year with one or more prescribed medications[8]. The assessment method for Recording of blood pressure, height, and weight was calculating the proportion of patients with one or more recordings in each of these fields[8]. Recording of Blood Pressure among Patients Requiring a Blood Pressure Measurement was examined by calculating the percentage of patients with one or more blood pressure recordings for patients with diabetes mellitus and patients with hypertension medications[8].

Correctness Domain: Two measures were developed by Terry et al. to examine correctness: Positive predictive value and unlikely combinations. Positive predictive value (PPV) was assessed by calculating PPVs for each of the six test conditions (Figure 1, see Appendix D for example calculation). Unlikely combinations of age and specific procedures, indicative of correctness, were assessed by calculating the percentage of patients 10 years or older who received a tetanus toxoid conjugate vaccination (usually only given to patients <10 years of age)[8]. Since lower scores on this measure indicate higher quality, the unlikely combinations score was subtracted from 100 (100- the percentage of patients 10 years or older who received a tetanus toxoid conjugate vaccination) so that all scores were directionally the same and could be combined.

Reference Standard vs. Billing Code 2X2 Table

	Reference standard criteria present	Reference standard criteria absent
Billing code present	True Positive	False Positive
Billing code absent	False Negative	True Negative

Sensitivity = True positive/ (True positive + False negative)
Positive Predictive Value = True positive / (True positive + False Positive)

Figure 1: Sensitivity and Positive Predictive Value Calculation

Currency/Timeliness Domain: Three measures were developed by Terry et al. to examine currency: timeliness of weight recording and timeliness of visit for pregnancy. Timeliness of weight recording for patients with obesity was assessed by calculating the percentage of obese patients in the database who had one or more weight recordings within a year of their last visit[8]. Timeliness of visit for pregnancy was assessed by calculating the percentage of patients with one or more visits within the two months following a positive pregnancy laboratory result[8]. Timeliness of blood pressure, height, and weight recording was assessed by calculating the proportion of patients with one or more recordings in each of these fields entered no more than one year prior to their last visit[8].

In order to calculate scores, the ideas underlying the assessment methods for this thesis had to be translated into R code. The R code was used to identify the relevant patient records within the data; for example, determining the patients who had hypertension based on the reference standard definition being used. To do this, the intermediate step of creating pseudocode was used. Pseudocode is a methodology frequently used in computer science in which the steps of an algorithm are written in plain language prior to being translated to the desired programming language[45]. Pseudocode has the advantage of being understandable, even for those who do not have programming knowledge. Pseudocode is a commonly used intermediary step when an idea or concept needs to be turned into code[45]. In these cases, and in this thesis, pseudocode allows clarification of principles and the breakdown of complex ideas into fully described discrete steps that can be executed.

3.8 Creation of Domain-Level Scores for Each Domain

Domain-level scores were created for each of the three domains of data quality (completeness, correctness, and currency/timeliness) according to each of the two methods, summation and arithmetic means. The creation of domain-level scores was done in order to be able to visualize how each domain contributes to the overall data quality score and to allow an additional point of examination and comparison for future users of this scoring system. Differences in data quality at the domain level have the potential to significantly impact the composite score and these differences can help to describe potential strengths and weaknesses of the data. The creation of domain-level scores also allowed comparison of the scoring methods within individual domains in addition to comparison of the overall score. In future use of the scoring system, the creation of individual domain scores will allow researchers to identify if domains contribute asymmetrically to overall data quality for a particular dataset. Identification of variation in data quality between domains will allow researchers to recognize potential data quality issues in a more specific way that may in turn allow targeted solutions to data quality problems.

To visualize the individual assessment method contributions to the domain scores, radar charts were created. The radar charts allow easy visual comparison of the assessment methods within the domain as they all have a theoretical maximum score of 100% and some variation is expected between them within domains.

For the summation method, the results of the 14 assessment methods contributing to the completeness domain score were summed, the seven assessment methods contributing to the correctness domain score were summed, and the five assessment methods contributing to the currency/timeliness domain score were summed to create one score for each domain; that is, the domain-level scores.

For the arithmetic mean method, the results of the 14 assessment methods contributing to the completeness domain score were added and the product was divided by 14, the seven assessment methods contributing to the correctness domain score were added and the product was divided by seven, and the five assessment methods contributing to the currency/timeliness domain score were added and the product was divided by five to create one score for each domain.

3.9 Creation of Composite Scores

The creation of the composite scores is diagrammed in Figure 2. For each of the scoring methods (summation and arithmetic mean), the domain-level scores were combined in the same manner as the individual assessment methods were combined. In the summation method, the three domain-level scores (calculated through summation) were summed to create the composite score. In the arithmetic mean method, the three domain-level scores (calculated by taking the arithmetic mean of the assessment methods) were summed and divided by three to create a composite score. While it is frequently not appropriate to take the arithmetic mean of a set of arithmetic means, in this case it is desirable to equally weight each of the three domains despite different numbers of assessment methods contributing to each of the three domains. This equal weighting is achieved by taking the arithmetic mean of the already calculated arithmetic means for each domain rather than taking the arithmetic mean of all of the assessment methods. The highest achievable score using the summation method would be 2600 and the highest achievable score using the arithmetic mean method would be 100.

To visualize the domain-level contributions to the composite scores, radar charts were created. The radar charts allow easy visual comparison of the domain-level scores within the composite scores.

To examine how the scoring systems differed, the domain-level scores derived from each method and the composite scores derived from each method were compared within and across the Development and Replication Groups. Differences in how much difference individual assessment methods contributed to domain scores and how much domain scores contributed to the composite score were compared. Further, the Development and Replication scores were compared to examine the reliability of the scoring method.

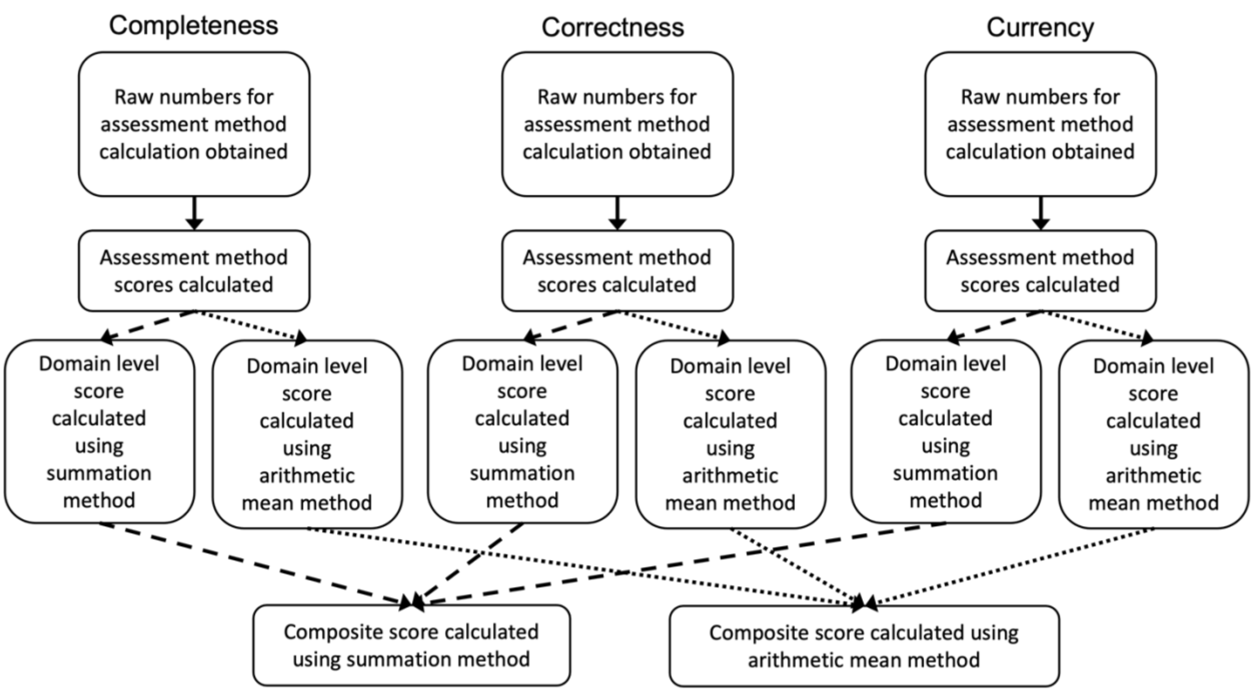

Figure 2: Flow diagram of composite score creation

3.10 Examination of Score Reliability

In order to determine the reliability of each of the scoring methods, the domain-level and composite scores obtained through the steps described in sections 3.7, 3.8, and 3.9 were calculated for the Replication Group (the second subgroup into which the data was divided). The scores derived from the Development Group were compared to those calculated for the Replication Group. Because the creation of the two groups was random and the dataset large, it was expected the scores calculated for the Development Group and for the Replication Group would be approximately equal. Any major differences between group scores within a single scoring system could be indicative of a lack of reliability in the scoring method.

Chapter 4

4 Results

4.1 Descriptive Analysis

The total dataset consisted of 1,839,101 patients. Patient demographics for the total dataset and each of the development and replication data subsets are described in Table 4.

Table 4: Patient Demographics (column %)

	Total Dataset	Development Group	Replication Group
Total number of Patients	1,839,101	919,551	919,550
Sex			
Males (n, %)	839,797 (45.6%)	419421 (45.6%)	420376 (45.7%)
Females (n, %)	995,649 (54.1%)	499818 (54.3%)	495831 (53.9%)
Missing Sex (n, %)	3,655 (0.2%)	312 (0.0%)	3343 (0.6%)
Patient Birth year			
1890-1899	542 (0.03%)	14 (0.0%)	528 (0.1%)
1900-1909	1,387 (0.1%)	263 (0.03%)	1,124 (0.1%)
1910-1919	9,274 (0.5%)	2,894 (0.3%)	6,380 (0.7%)
1920-1929	46,685 (2.5%)	19,357 (2.1%)	27,328 (3.0%)
1930-1939	87,853 (4.8%)	40,170 (4.4%)	47,683(5.2%)
1940-1949	149,119 (8.1%)	70,386 (7.6%)	78,733 (8.6%)
1950-1959	216,850 (11.8%)	109,991 (12.0%)	106,859 (11.6%)
1960-1969	240,553 (13.1%)	126,717 (13.8%)	113,836 (12.4%)
1970-1979	237,340 (12.9%)	127,023 (13.8%)	110,317 (12.0%)
1980-1989	257,210 (14.0%)	133,626 (14.5%)	123,584 (13.4%)
1990-1999	214,690 (11.7%)	107,848 (11.7%)	106,842 (11.6%)
2000-2009	191,935 (10.4%)	9,6031 (10.4%)	95,904 (10.4%)
2010-2019	184,177 (10.0%)	84,240 (9.2%)	99,937 (10.9%)
Missing Birth Year	1,486 (0.1%)	991 (0.1%)	495 (0.1%)
Habitation			
Urban (n, %)	1,379,755 (75.0%)	759,489 (82.6%)	620,266 (67.4%)
Rural (n, %)	272,613 (14.8%)	107,899 (11.7%)	164,714 (17.9%)
Missing Location Code (n, %)	186,733 (10.1%)	52,163 (5.7%)	134,570 (14.6%)

The t-test for patient birth year indicated the Development Group (M=1976, SD= 23.7) was born later on average than the Replication Group (M=1975, SD=25.6), t-value with >500 degrees of freedom was 27.2 with a p-value of <0.01. The effect size estimate was very small, Cohen's d = 0.04 (95% CI 0.04, 0.04) indicating that, while statistically significant, there was no important difference in birth year between the groups. The Chi square test for patient sex indicated a difference between the Development and Replication Groups, χ^2 (2, N=1839101) = 2530.6, $p<0.01$. However, the effect size estimate was very small, Cramer's V = 0.04, indicating the effect was negligible and therefore the difference was negligible for this analysis. The Chi square test of patient location indicated a difference between the Development and Replication Groups, χ^2 (2, N=1839101) = 62256, $p<0.01$. However, the effect size estimate was small, Cramer's V = 0.18, indicating the difference was unimportant for this analysis. Therefore, the Development and Replication Groups were similar to each other in distributions of age, sex, and location. The groups were also similar in proportions of missing sex and birth year. There was a slight difference in the percentage of missing location codes, as indicated by the marginally larger effect size in the location Chi square test of location; this variable was not used in the data quality scoring. Some patients have birth years that would make them over 120 years old. These patient records exist in the database because it is a longitudinal database encompassing patient records from multiple years. The records used in this thesis are records that were added to the CPCSSN database between January 1st 2015 and December 31st 2019. Some of the patients included in the analyses in this thesis are inevitably deceased but their date of death and even whether they are deceased are not always included in their EMR. Whether patients are deceased did not affect any of the analyses conducted in this thesis.

There were 1,497 care providers in the database. These care providers used 11 different EMR systems: Accuro[46], Healthquest[47], InputHealth[48], Intrahealth[49], Med Access[50], Medesync[51], OSCAR[52], P&P[53], Practice Solutions[54], Purkinje[55], Wolf[56]. Of these EMR systems, Practice

Solutions was the most common, used by 26% of the providers, and Healthquest was the least prevalent with less than 1% of providers using that EMR.

4.2 Pseudocode

Pseudocode was created for each of the 14 assessment methods by going through each assessment method and elaborating each of the steps required to identify the groups of interest given the data format. The format of data contained in each column, e.g. text or numeric, is not referred to in the pseudocode. Most of the columns used were in a free text format, nvarchar, though dates were in a date format, YYY-MM-DD, and patient identifiers were numeric. Table 5 reports the pseudocode for the 14 assessment methods. This pseudocode was used as the intermediary step to creating R code but is a deliverable in and of itself, in that, it can be shared with other researchers to aid in the consistent creation of these metrics for future research projects.

Table 5: Pseudocode for identifying patients relevant to assessment methods

Assessment Method	Pseudocode
Completeness (Sensitivity): Sensitivity value for diabetes mellitus (test condition definition vs. billing code)	Reference standard positives for diabetes: Number of patients (unique patient IDs) with: “DIABETES MELLITUS” in the ‘DiagnosisText_calc’ column of the ‘HealthCondition’ table OR “250” or “250.x” (x could be any number)) in the ‘DiagnosisCode_calc’ column of the ‘HealthCondition’ table OR one of [medications omitted. See Appendix C] in the ‘Name_calc’ column of the ‘Medication’ table, OR >=2 (at least 2) of: [medications omitted. See Appendix C] In the ‘Name_calc’ column of the ‘Medication’ table OR 1 or more of: [medications omitted. See Appendix C] In the ‘Name_calc’ column of the ‘Medication’ table OR 2 or more of: • “Fasting Blood Glucose(FBG)” in the ‘Name_calc’ column of the ‘Lab’ table and a value of 7.0 or greater in the corresponding ‘TestResult_calc’ column • “Glucose [Moles/volume] in Serum or Plasma” in the ‘Name_calc’ column of the ‘Lab’ table and a value of 11.1 or greater in the corresponding ‘TestResult_calc’ column • “Glucose Tolerance Test (GTT)” in the ‘Name_calc’ column of the ‘Lab’ table and a value of 11.1 or greater in the corresponding ‘TestResult_calc’ column For all above with 3 or more months difference between the dates in the ‘PerformedDate’ column (YYYY-MM-DD) of the ‘Lab’ table corresponding to each result OR “Hemoglobin A1c (HbA1c)” in the ‘Name_calc’ column of the ‘Lab’ table and a value of 6.5 or greater in the corresponding ‘TestResult_calc’ column of the ‘Lab’ table.

Assessment Method	Pseudocode
	Of those patients (unique patient IDs) who meet at least 1 of the above five criteria, how many have the code “250” or “250.x” (x could be any number) in the ‘DiagnosisCode_calc’ column of the ‘Billing’ table
Completeness (Sensitivity): Sensitivity value for hypertension (test condition definition vs. billing code)	Reference standard positives for hypertension: Number of patients (unique patient IDs) with: a difference of 18 or greater between the year in the ‘BirthYear’ column of the ‘Patient’ table and 2019 with: “HYPERTENSION” in the ‘DiagnosisText_calc’ column of the ‘HealthCondition’ table OR “401” or “401.x” (x could be any number) in the ‘DiagnosisCode_calc’ column of the ‘HealthCondition’ table OR 3 or more (entries of “sBP (mmHg)” in the ‘Exam1’ column of the ‘Exam’ table and a value greater than or equal to 140 in the corresponding ‘Result1_calc’ column of the ‘Exam’ table) OR 3 or more (entries of “dBP (mmHg)” in the ‘Exam2’ column of the ‘Exam’ table and a value greater than or equal to 90 in the corresponding ‘Result2_calc’ column of the ‘Exam’ table) OR 1 or more of: [medications omitted. See Appendix C] In the ‘Name_calc’ column of the ‘Medication’ table OR 2 or more of [medications omitted. See Appendix C] In the ‘Name_calc’ column of the ‘Medication’ table Of those patients (unique patient IDs) who meet at least 1 of the above criteria how many have “401” or “401.x” (x could be any number) in the ‘DiagnosisCode_calc’ column of the ‘Billing’ table
Completeness (Sensitivity): Sensitivity value for hypothyroidism (test condition definition vs. billing code)	Reference standard positives for hypothyroidism: Number of patients (unique patient IDs) with: “HYPOTHYROIDISM” in the ‘DiagnosisText_calc’ column of the ‘HealthCondition’ table OR “244” or “244.x” (x could be any number) in the ‘DiagnosisCode_calc’ column of the ‘HealthCondition’ table OR

Assessment Method	Pseudocode
	1 or more of: [medications omitted. See Appendix C] In the 'Name_calc' column of the 'Medication' table OR "Thyroid-Stimulating Hormone (TSH)" in the 'Name_calc' column of the 'Lab' table and a value greater than 10.0 in the corresponding 'TestResult_calc' column of the 'Lab' table OR "TSH" in the 'Name_calc' column of the 'Lab' table and a value greater than 10.0 in the corresponding 'TestResult_calc' column of the 'Lab' table Of those patients (unique patient IDs)who meet at least 1 of the criteria how many have "244" or "244.x" (x could be any number) in the 'DiagnosisCode_calc' column of the 'Billing' table
Completeness (Sensitivity): Sensitivity value for asthma (test condition definition vs. billing code)	Reference standard positives for asthma: Number of patients (unique patient IDs) with: "ASTHMA" in the 'DiagnosisText_calc' column of the 'HealthCondition' table OR "493" or "493.x" (x could be any number) in the 'DiagnosisCode_calc' column of the 'HealthCondition' table) AND 1 or more of: [medications omitted. See Appendix C] In the 'Name_calc' column of the 'Medication' table) OR (1 or more of: [medications omitted. See Appendix C] In the 'Name_calc' column of the 'Medication' table AND 1 or more of: [medications omitted. See Appendix C] In the 'Name_calc' column of the 'Medication' table) Of those patients (unique patient IDs) who meet at least 1 of the criteria how many have "493" or "493.x" (x could be any number) in the 'DiagnosisCode_calc' column of the 'Billing' table
Completeness (Sensitivity): Sensitivity value for obesity (test condition definition vs. billing code)	Reference Standard Positives: Number of Patients (unique patient IDs) with: a difference greater than 18 between the value in the 'BirthYear' column in the 'Patient' table and 2019 with: "OBESE" or "OBESITY" in the 'DiagnosisText_calc' column of the 'HealthCondition' table OR

Assessment Method	Pseudocode
	"278" or "278.x" (x could be any number) in the 'DiagnosisCode_calc' column of the 'HealthCondition' table OR "BMI (kg/m^2)" in the 'Exam1' column of the 'Exam' table and a value of 30 or greater in the corresponding 'Result1_calc' column of the 'Exam table Of those patients (unique patient IDs)who meet at least 1 of the criteria how many have "278" or "278.x" (x could be any number) in the 'DiagnosisCode_calc' column of the 'Billing' table
Completeness (Sensitivity): Sensitivity value for urinary tract infection (test condition definition vs. billing code)	Reference standard positives for urinary tract infection: Number of patients (unique patient IDs) with: "URINARY TRACT INFECTION" or "UTI" in the 'DiagnosisText_calc' column of the 'HealthCondition' table OR "595" or "595.x" (x could be any number) in the 'DiagnosisCode_calc' column of the 'HealthCondition' table AND 1 or more of: [medications omitted. See Appendix C] In the 'Name_calc' column of the 'Medication' table Of those patients (unique patient IDs)who the criteria how many have "595" or "595.x" in the 'DiagnosisCode_calc' column of the 'Billing' table
Completeness (Consistency of capture): % of patients with 1 or more problem list entries	Number of unique entries in the 'Patient_ID' column exist in the 'HealthCondition' table
Completeness (Consistency of capture): % of patients with 1 or more allergy record entries	Number of unique entries in the 'Patient_ID' column exist in the 'AllergyIntolerance' table
Completeness (Consistency of capture): % of patients who visited in past year with 1 or more prescribed meds	Identify unique entries in 'Patient_ID' column of 'Encounter' table that also have an entry in the 'DateCreated' column of the 'Encounter' table with a year of 2019 (data format of 'DateCreated' column is YYYY-MM-DD). Report the number of patients meeting that criteria. Of these patients (the unique Patient_ID entries meeting the above criteria) what this the number that also appear in the 'Patient_ID' column of the 'Medication' table
Completeness (Recoding of BP, height, and weight):	Number of unique entries in the 'Patient_ID' column of the of the 'Patient' table that have at least 1 corresponding entry of "sBP (mmHg)" in the 'Exam1' column of the 'Exam' table

Assessment Method	Pseudocode
% of patients with 1 or more blood pressure recordings for patients 18 + years	
Completeness (Recoding of BP, height, and weight): % of patients with 1 or more height recordings	Number of unique entries in the ‘Patient_ID’ column of the of the ‘Patient’ table that have at least 1 corresponding entry of “Height (cm)” in the ‘Exam1’ column of the ‘Exam’ table
Completeness (Recoding of BP, height, and weight): % of patients with 1 or more weight recording	Number of unique entries in the ‘Patient_ID’ column of the of the ‘Patient’ table that have at least 1 corresponding entry of “Weight (kg)” in the ‘Exam1’ column of the ‘Exam’ table
Completeness (Recording of BP among patients requiring a BP measurement): % of patients with diabetes mellitus, with 1 or more blood pressure recordings	Number of patients (unique patient IDs) with: “DIABETES MELLITUS” in the ‘DiagnosisText_calc’ column of the ‘HealthCondition’ table OR “250” or “250.x” (x could be any number)) in the ‘DiagnosisCode_calc’ column of the ‘HealthCondition’ table OR one of [medications omitted. See Appendix C] in the ‘Name_calc’ column of the ‘Medication’ table, OR >=2 (at least 2) of: [medications omitted. See Appendix C] In the ‘Name_calc’ column of the ‘Medication’ table OR 1 or more of: [medications omitted. See Appendix C] In the Name_calc’ column of the ‘Medication’ table OR 2 or more of: • “Fasting Blood Glucose(FBG)” in the ‘Name_calc’ column of the ‘Lab’ table and a value of 7.0 or greater in the corresponding ‘TestResult_calc’ column • “Glucose [Moles/volume] in Serum or Plasma” in the ‘Name_calc’ column of the ‘Lab’ table and a value of 11.1 or greater in the corresponding ‘TestResult_calc’ column • “Glucose Tolerance Test (GTT)” in the ‘Name_calc’ column of the ‘Lab’ table and a value of 11.1 or greater in the corresponding ‘TestResult_calc’ column

Assessment Method	Pseudocode
	For all above with 3 or more months difference between the dates in the ‘PerformedDate’ column (YYYY-MM-DD) of the ‘Lab’ table corresponding to each result OR “Hemoglobin A1c (HbA1c)” in the ‘Name_calc’ column of the ‘Lab’ table and a value of 6.5 or greater in the corresponding ‘TestResult_calc’ column of the ‘Lab’ table. Number of patients (unique patient IDs) who meet the above criteria who have a corresponding entry of “sBP (mmHg)” in the ‘Exam1’ column of the ‘Exam’ table
Completeness (Recording of BP among patients requiring a BP measurement): % of patients with hypertension medications (2 or more oral anti- hypertensives, or 1 or more diuretics) with 1 or more blood pressure recordings	Number of patients (unique patient IDs) with a difference of 18 or greater between the year in the ‘BirthYear’ column of the ‘Patient’ table and 2019 with: “HYPERTENSION” in the ‘DiagnosisText_calc’ column of the ‘HealthCondition’ table OR “401” or “401.x” (x could be any number) in the ‘DiagnosisCode_calc’ column of the ‘HealthCondition’ table OR 3 or more (entries of “sBP (mmHg)” in the ‘Exam1’ column of the ‘Exam’ table and a value greater than or equal to 140 in the corresponding ‘Result1_calc’ column of the ‘Exam’ table) OR 3 or more (entries of “dBP (mmHg)” in the ‘Exam2’ column of the ‘Exam’ table and a value greater than or equal to 90 in the corresponding ‘Result2_calc’ column of the ‘Exam’ table) OR 1 or more of: [medications omitted. See Appendix C] In the ‘Name_calc’ column of the ‘Medication’ table OR 2 or more of [medications omitted. See Appendix C] In the ‘Name_calc’ column of the ‘Medication’ table Number of patients (unique patient IDs) who meet the above criteria who have a corresponding entry of “sBP (mmHg)” in the ‘Exam1’ column of the ‘Exam’ table
Correctness (Positive predictive value):	Same code as diabetes mellitus sensitivity

Assessment Method	Pseudocode
Positive predictive value for diabetes mellitus (test condition definition vs. billing code)	
Correctness (Positive predictive value): Positive predictive value for hypertension (test condition definition vs. billing code)	Same code as hypertension sensitivity
Correctness (Positive predictive value): Positive predictive value for hypothyroidism (test condition definition vs. billing code)	Same code as hypothyroidism sensitivity
Correctness (Positive predictive value): Positive predictive value for asthma (test condition definition vs. billing code)	Same code as asthma sensitivity
Correctness (Positive predictive value): Positive predictive value for obesity (test condition definition vs. billing code)	Same code as obesity sensitivity
Correctness (Positive predictive value): Positive predictive value for urinary tract infection (test condition definition vs. billing code)	Same code as urinary tract infection sensitivity

Assessment Method	Pseudocode
Correctness (unlikely combinations of age and specific procedures): % of patients ≥10 yrs. of age with a tetanus toxoid conjugate vaccine (diphtheria, haemophilus B, pertussis, polio, and tetanus)	Number of patients (unique patient IDs) who have an entry of “Diphtheria-Pertussis-Poliomyelitis-Tetanus” or “Dipht-Hemoph Influen B-Pertuss-Polio-Tetanus” or “Diphtheria-Poliomyelitis-Tetanus” in the ‘Name_calc’ column of the ‘Vaccine’ table where the corresponding year in the date in the ‘DateCreated’ column of the ‘Vaccine’ table (YYY-MM-DD format) is more than 10 years after the entry in the ‘BirthYear’ column of the ‘Patient’ table
Currency (Timeliness of weight recordings for patients with obesity): % of obese patients with 1 or more weight recordings within 1 year of last visit in recorded in the database	Number of Patients (unique patient IDs) with a difference greater than 18 between the value in the ‘BirthYear’ column in the ‘Patient’ table and 2019 with: “OBESE” or “OBESITY” in the ‘DiagnosisText_calc’ column of the ‘HealthCondition’ table OR “278” or “278.x” (x could be any number) in the ‘DiagnosisCode_calc’ column of the ‘HealthCondition’ table OR “BMI (kg/m^2)” in the ‘Exam1’ column of the ‘Exam’ table and a value of 30 or greater in the corresponding ‘Result1_calc’ column of the ‘Exam table Number of patients (unique patient IDs) that meet the above criteria with one or more entries of “Weight (kg)” in the ‘Exam1’ column of the ‘Exam’ table with the where the difference between the date in the ‘DateCreated’ column, corresponding to the appearance of “Weight (kg)” in the ‘Exam1’ column and the most recent date in the ‘DateCreated’ column of the ‘Encounter’ Table for each identified unique patient ID is less than 1 year.
Currency (timeliness of visit for antenatal care): % of patients with a positive pregnancy laboratory test result and 1 or more visits within two months of the result	Number of patients (unique patient IDs) with an entry “Choriogonadotropin.beta subunit (pregnancy test) [Presence] in Serum or Plasma” in the ‘Name_calc’ column of the ‘Lab’ table and a value of 25 or greater in the corresponding ‘TestResult_calc’ OR an entry of “Choriogonadotropin Beta Subunit in Serum or Plasma (BHCG)” in the ‘Name_calc’ column of the ‘Lab’ table and a value of 25 or greater in the corresponding ‘TestResult_calc’ column Of patients (unique patient IDs) who meet the above criteria what is the number who have at least one entry in the ‘DateCreated’ column of the ‘Encounter’ table where there is a difference of less

Assessment Method	Pseudocode
	than 2 months between the month and year in the ‘DateCreated’ column of the ‘Encounter’ table and the month and year in the ‘DateCreated’ column of the ‘Lab’ corresponding to the result above
Currency (timeliness of bp, height, and weight recordings): % of 18yr.+ patients with 1 or more blood pressure values recorded ≤ one year prior to their last visit in the database	Number of patients (unique patient IDs) with a difference of 18 or greater between the year in the ‘BirthYear’ column of the ‘Patient’ table and 2019 with one or more entries of “sBP (mmHg)” in the ‘Exam1’ column of the ‘Exam’ table where the date in the ‘DateCreated’ column of the ‘Exam’ table corresponding with the ‘Exam1’ entry is less than or equal to the most recent date corresponding with the patient ID in the ‘DateCreated’ column of the ‘Encounter’ table
Currency (timeliness of bp, height, and weight recordings): %of patients with 1 or more height values recorded ≤ one year prior to their last visit in the database	Number of patients (unique patient IDs) with one or more entries of “Height (cm)” in the ‘Exam1’ column of the ‘Exam’ table where the date in the ‘DateCreated’ column of the ‘Exam’ table corresponding with the ‘Exam1’ entry is less than or equal to the most recent date corresponding with the patient ID in the ‘DateCreated’ column of the ‘Encounter’ table
Currency (timeliness of bp, height, and weight recordings): % of patients with 1 or more weight values recorded ≤ one year prior to their last visit in the database for	Number of patients (unique patient IDs) with one or more entries of “Weight (kg)” in the ‘Exam1’ column of the ‘Exam’ table where the date in the ‘DateCreated’ column of the ‘Exam’ table corresponding with the ‘Exam1’ entry is less than or equal to the most recent date corresponding with the patient ID in the ‘DateCreated’ column of the ‘Encounter’ table

4.3 Creation of Domain and Composite Scores

This section reports the creation of the domain-level and composite scores. These scores were calculated using the subset of the data referred to as the Development Group. The calculated scores for the assessment methods and domains are reported as well as the composite score.

4.3.1 Completeness Domain

A summary of the scores on each of the assessment methods of completeness can be found in Table 6 grouped under their corresponding Measure. The scores on the assessment methods of completeness were generally mid-range with a few deviations. In comparison to the other sensitivity values, the sensitivity value calculated for obesity was particularly low (4.4%). The scores for the percentages of patients who required blood pressure recordings due to a diagnosis of diabetes or hypertension and who received them were particularly high (89.7% and 90.2% respectively).

Table 6: Development Group Completeness Scores

Measure	Assessment method	Development Group Score
Sensitivity	Sensitivity value for diabetes mellitus (test condition definition vs. billing code)	17.7%
	Sensitivity value for hypertension (test condition definition vs. billing code)	31.9%
	Sensitivity value for hypothyroidism (test condition definition vs. billing code)	26.4%
	Sensitivity value for asthma (test condition definition vs. billing code)	34.5%
	Sensitivity value for obesity (test condition definition vs. billing code)	4.4%
	Sensitivity value for urinary tract infection (test condition definition vs. billing code)	17.8%
Consistency of capture	% of patients with 1 or more problem list entries	62.7%
	% of patients with 1 or more allergy record entries	40.6%
	% of patients who visited in past year with 1 or more prescribed meds	92.0%

Measure	Assessment method	Development Group Score
Recording of blood pressure, height, and weight	% of patients with 1 or more blood pressure recordings for patients 18 + years	70.6%
	% of patients with 1 or more height recordings	61.7%
	% of patients with 1 or more weight recording	66.7%
Recording of blood pressure among patients requiring a blood pressure measurement	% of patients with diabetes mellitus, with 1 or more blood pressure recordings	89.7%
	% of patients with hypertension medications (2 or more oral anti- hypertensives, or 1 or more diuretics) with 1 or more blood pressure recordings	90.2%

Using the summation method, the combined completeness domain score for the Development Group was calculated to be 706.8 out of the theoretical maximum of 1400. Using the arithmetic mean method, the combined completeness score for the Development Group was calculated to be 50.5 out of the theoretical maximum of 100. The relative contributions of each assessment method to the completeness domain score are illustrated in Figure 3. Assessment methods with higher scores, such as 'the percentage of patients who visited in the past year with one or more prescribed medications', contributed relatively more to the score than those with lower scores, such as 'the sensitivity value for obesity'.

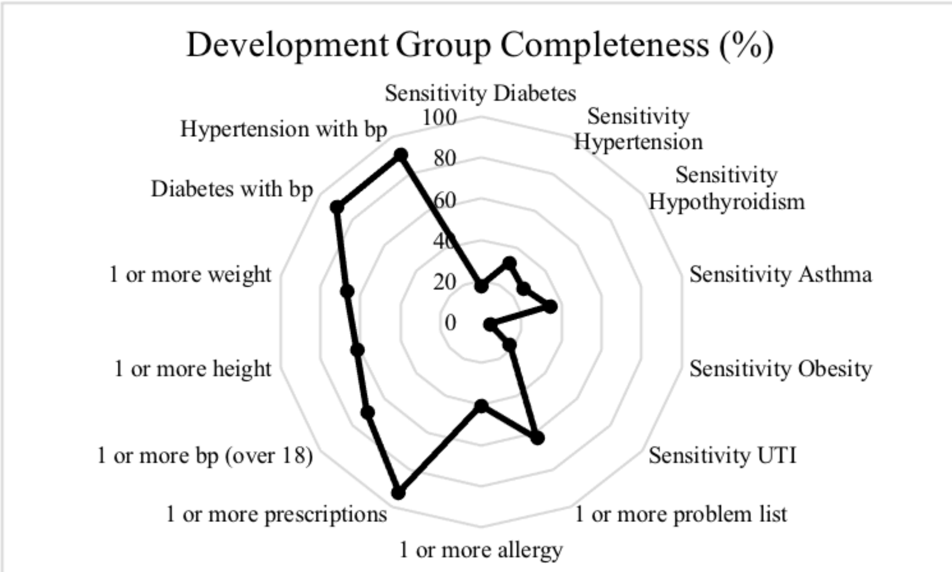

Figure 3: Completeness component scores – Development Group

4.3.2 Correctness Domain

A summary of the scores on each of the assessment methods of completeness can be found in Table 7. The scores on the assessment methods of correctness were generally mid-range to high. Two exceptions were the positive predictive value for urinary tract infections 1.7% and hypothyroidism at 18.6%.

Table 7: Development Group Correctness Scores

Measure	Assessment Method	Development Group Score
Positive predictive value	Positive predictive value for diabetes mellitus (test condition definition vs. billing code)	84.7%
	Positive predictive value for hypertension (test condition definition vs. billing code)	89.4%
	Positive predictive value for hypothyroidism (test condition definition vs. billing code)	18.6%
	Positive predictive value for asthma (test condition definition vs. billing code)	43.2%
	Positive predictive value for obesity (test condition definition vs. billing code)	41.7%
	Positive predictive value for urinary tract infection (test condition definition vs. billing code)	1.7%

Measure	Assessment Method	Development Group Score
Unlikely combinations of age & specific procedures	100 minus % of patients ≥10 yrs. of age with a tetanus toxoid conjugate vaccine (diphtheria, haemophilus B, pertussis, polio, and tetanus)	93.2%

Using the summation method, the combined correctness domain score for the Development Group was calculated to be 372.4 out of the theoretical maximum of 700. Using the arithmetic mean method, the combined correctness score for the Development Group was calculated to be 53.2 out of the theoretical maximum of 100. The relative contribution of each assessment method to the correctness domain score is illustrated in Figure 4. Assessment methods with higher scores, such as ‘unlikely combinations of age & specific procedures’, contributed relatively more to the score than those with lower scores, such as ‘positive predictive value for urinary tract infection’.

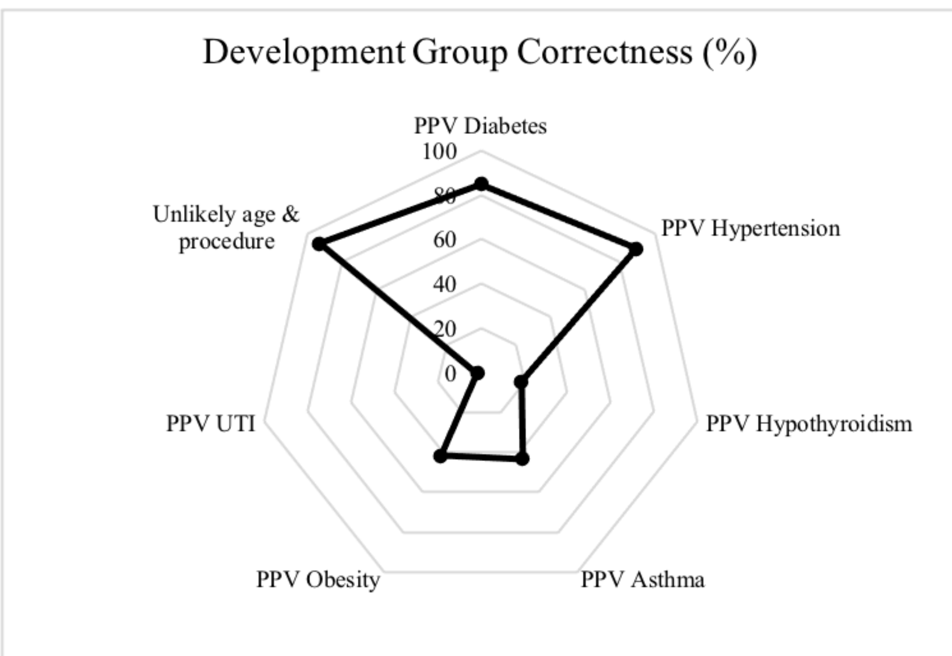

Figure 4: Development group correctness component scores

4.3.3 Currency Domain

A summary of the scores on each of the assessment methods of completeness can be found in Table 8. The scores on the assessment methods of currency were high to very high. The highest score was “timeliness of weight recordings for patients with obesity” which was 99.5%. In general, this domain had the highest scores of the three domains.

Table 8: Development Group Currency Scores

Measure	Assessment Method	Development Group Score
Timeliness of weight recordings for patients with obesity	% of obese patients with 1 or more weight recordings within 1 year of last visit in recorded in the database	99.5%
Timeliness of visit for antenatal care	% of patients with a positive pregnancy laboratory test result and 1 or more visits within two months of the result	95.8%
Timeliness of blood pressure, height, and weight recordings	% of 18yr.+ patients with 1 or more blood pressure values recorded ≤ one year prior to their last visit in the database	67.6%
	%of patients with 1 or more height values recorded ≤ one year prior to their last visit in the database	59.4%
	% of patients with 1 or more weight values recorded ≤ one year prior to their last visit in the database for	64.0%

Using the summation method, the combined currency domain score for the Development Group was calculated to be 389.3 out of a theoretical maximum of 500. Using the arithmetic mean method, the combined currency score for the Development Group was calculated to be 77.2 out of a theoretical maximum of 100. The relative contributions of each assessment method to the currency domain score in each of the four cases is illustrated in Figure 5. Assessment methods with higher scores, such as ‘Timeliness of weight recordings for patients with obesity’, contributed relatively more to the score than those with lower scores, such as ‘timeliness of height recordings’.

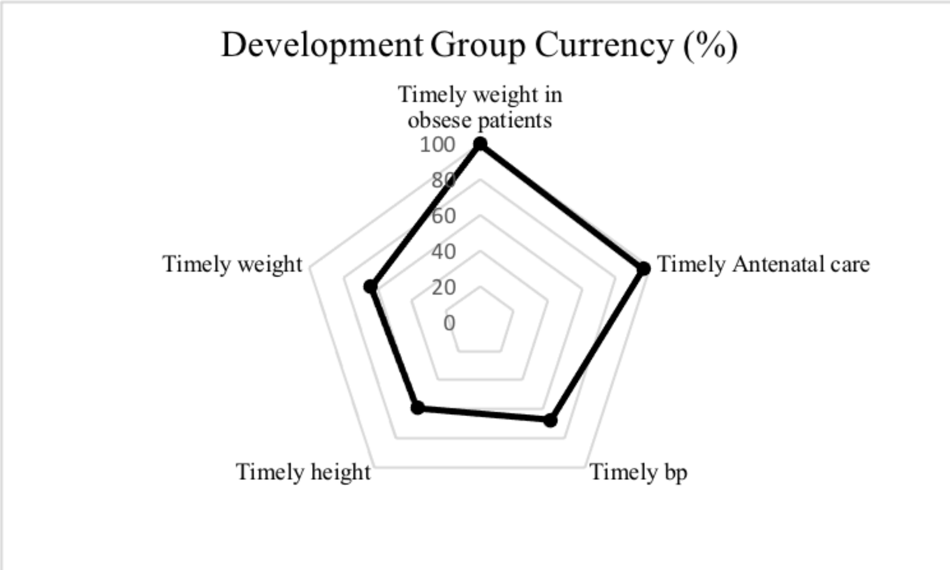

Figure 5: Development group currency component scores

4.3.4 Composite Scores

Two composite scores were calculated for the Development Group, one for each of the two methods (summation and arithmetic mean). Using the summation method, the Development Group composite score was calculated as 1465.5 out of a theoretical maximum of 2600. Using the arithmetic mean method, the Development Group composite score was calculated as 60.3 out of a theoretical maximum of 100. The relative contributions of the domain scores to the composite score is illustrated in Figures 6 and 7. In the summation score figure (Figure 6), it is important to note that because of the differential numbers of assessment methods contributing to each of the domains (completeness: 14 assessment methods, maximum score of 1400; correctness: 7 assessment methods, maximum score of 700; currency: 5 assessment methods, maximum score of 500) there are different possible maximum scores on each axis.

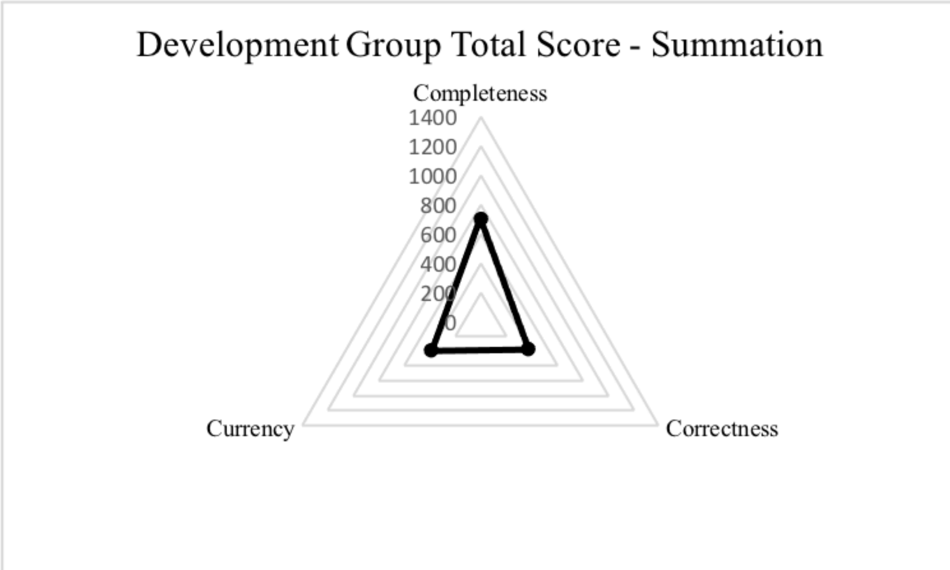

Figure 6: Development group total component scores - Summation method

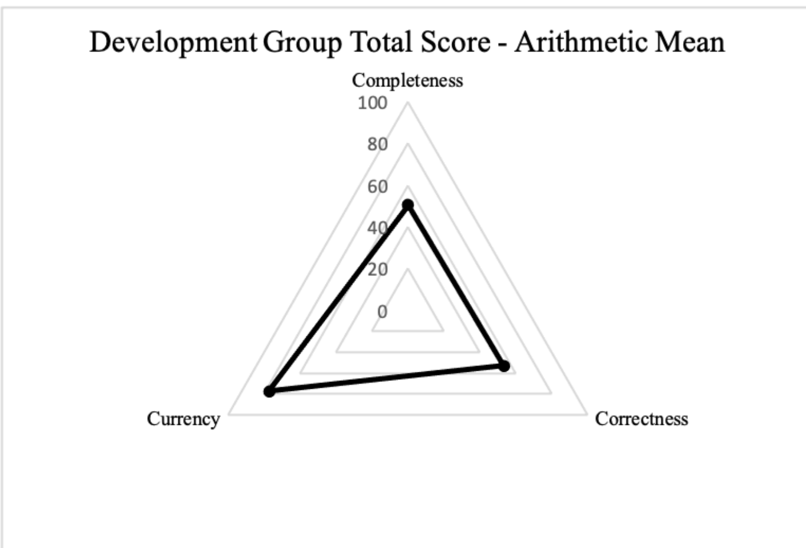

Figure 7: Development group total component scores - Arithmetic mean method

4.4 Replication of Domain and Composite Scores

4.4.1 Completeness

The scores on the assessment methods of completeness were generally mid-range to high. The lowest scores were in the calculated sensitivity values for obesity and urinary tract infections (10.0% and 12.3% respectively). The highest score was the percentage of patients with a prescription on file who visited within the past year (96.6%). A summary of the scores on each of the assessment methods of completeness can be found in Table 9.

Table 9: Replication Group Completeness Scores

Measure	Assessment Method	Replication Group Score
Sensitivity	Sensitivity value for diabetes mellitus (test condition definition vs. billing code)	20.4%
	Sensitivity value for hypertension (test condition definition vs. billing code)	37.6%
	Sensitivity value for hypothyroidism (test condition definition vs. billing code)	18.0%
	Sensitivity value for asthma (test condition definition vs. billing code)	28.8%
	Sensitivity value for obesity (test condition definition vs. billing code)	10.0%
	Sensitivity value for urinary tract infection (test condition definition vs. billing code)	12.3%
Consistency of capture	% of patients with 1 or more problem list entries	58.2%
	% of patients with 1 or more allergy record entries	45.5%
	% of patients who visited in past year with 1 or more prescribed meds	96.6%
Recording of blood pressure, height, and weight	% of patients with 1 or more blood pressure recordings for patients 18 + years	77.5%
	% of patients with 1 or more height recordings	55.3%
	% of patients with 1 or more weight recording	61.9%
Recording of blood pressure among patients requiring a blood pressure measurement	% of patients with diabetes mellitus, with 1 or more blood pressure recordings	93.7%
	% of patients with hypertension medications (2 or more oral anti- hypertensives, or 1 or more diuretics) with 1 or more blood pressure recordings	88.3%

Using the summation method, the combined completeness domain score for the Replication Group was 704.4 out of the theoretical maximum of 1400. Using the arithmetic mean method, the combined completeness score for the Replication Group was 50.3 out of a theoretical maximum of 100. The relative contribution of each assessment method to the completeness domain score is illustrated in Figure 8.

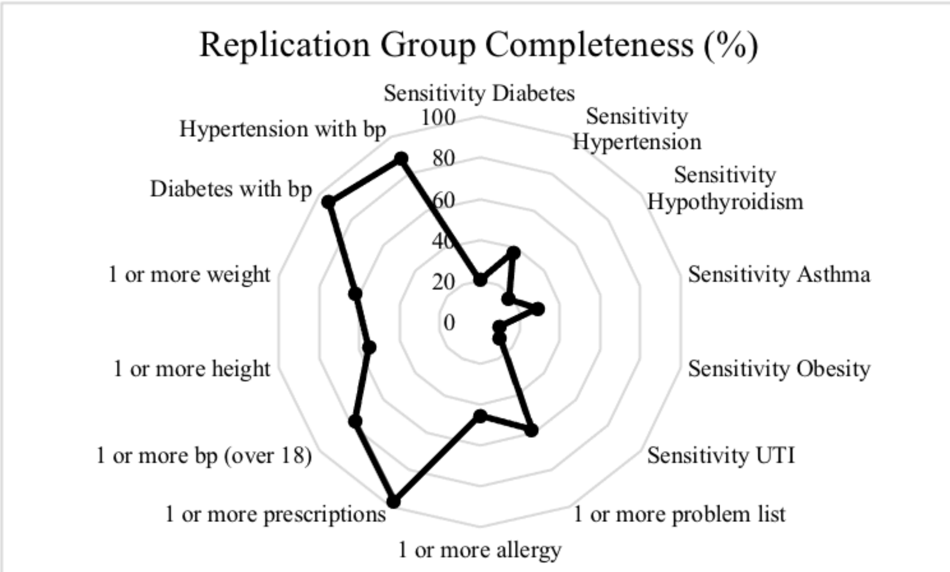

Figure 8: Replication group completeness component scores

4.4.2 Correctness

The scores on the assessment methods of correctness were mixed. There were several high scores, including the positive predictive value for diabetes (90.5%) and the unlikely combinations of age and procedure, calculated as the percentage of unlikely age specific procedures recorded subtracted from 100 (95.8%). However, the positive predictive value calculated for urinary tract infections was very low (3.1%). A summary of the scores on each of the assessment methods of completeness can be found in Table 10.

Table 10: Replication Group Correctness Scores

Measure	Assessment Method	Replication Group Score
Positive predictive value	Positive predictive value for diabetes mellitus (test condition definition vs. billing code)	90.5%
	Positive predictive value for hypertension (test condition definition vs. billing code)	87.4%
	Positive predictive value for hypothyroidism (test condition definition vs. billing code)	14.0%
	Positive predictive value for asthma (test condition definition vs. billing code)	36.2%

Measure	Assessment Method	Replication Group Score
	Positive predictive value for obesity (test condition definition vs. billing code)	47.9%
	Positive predictive value for urinary tract infection (test condition definition vs. billing code)	3.1%
Unlikely combinations of age & specific procedures	100- % of patients ≥10 yrs. of age with a tetanus toxoid conjugate vaccine (diphtheria, haemophilus B, pertussis, polio, and tetanus)	95.8%

Using the summation method, the combined correctness domain score for the Replication Group was calculated to be 374.7. Using the arithmetic mean method, the combined correctness score for the Replication Group was calculated to be 53.5. The theoretical maximum scores are 700 and 100 respectively. The relative contribution of each assessment method to the correctness domain score is illustrated in Figure 9.

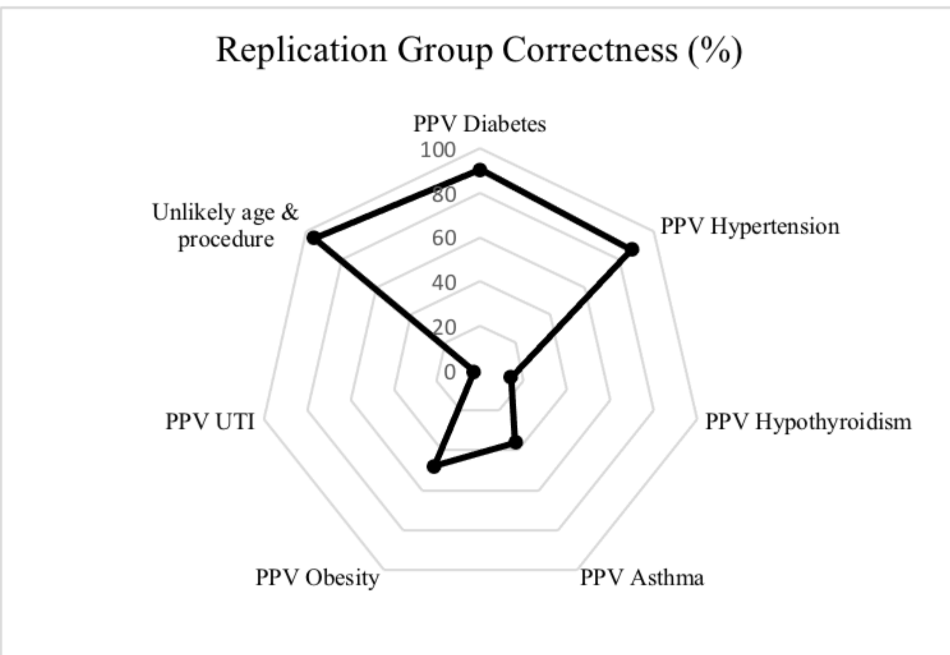

Figure 9: Replication group correctness component scores

4.4.3 Currency

The scores on the assessment methods of currency were generally high. The highest score was for timeliness of weight recordings in patients with obesity (97.0%). A summary of the scores on each of the assessment methods of completeness can be found in Table 11.

Table 11: Replication Group Currency Scores

Measure	Assessment Method	Replication Group Score
Timeliness of weight recordings for patients with obesity	% of obese patients with 1 or more weight recordings within 1 year of last visit in recorded in the database	97.0%
Timeliness of visit for antenatal care	% of patients with a positive pregnancy laboratory test result and 1 or more visits within two months of the result	93.3%
Timeliness of blood pressure, height, and weight recordings	% of 18yr.+ patients with 1 or more blood pressure values recorded ≤ one year prior to their last visit in the database	62.4%
	%of patients with 1 or more height values recorded ≤ one year prior to their last visit in the database	65.0%
	% of patients with 1 or more weight values recorded ≤ one year prior to their last visit in the database for	57.8%

Using the summation method, the combined currency domain score for the Replication Group was calculated to be 375.5. Using the arithmetic mean method, the combined currency score for the Replication Group was calculated to be 75.1. The theoretical maximum scores are 500 and 100 respectively. The relative contribution of each assessment method to the currency domain score is illustrated in Figure 10.

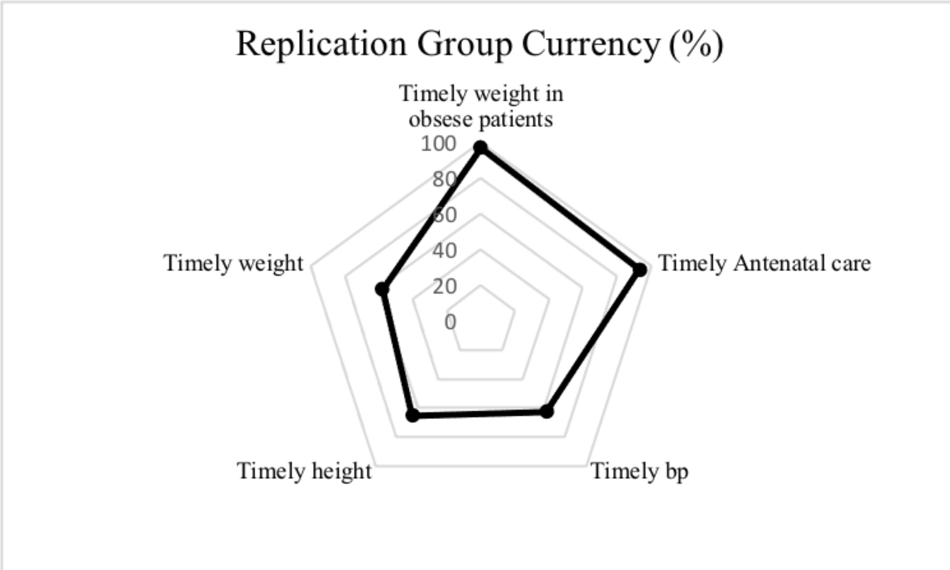

Figure 10: Replication group currency component scores

4.4.4 Composite Scores

Two composite scores were calculated, one for each of the two methods (summation and arithmetic mean). Using the summation method, the Replication Group composite score was calculated as 1454.6 out of a theoretical maximum of 2600. Using the arithmetic mean method, the Replication Group composite score was calculated as 59.6 out of a theoretical maximum of 100. The relative contributions of the domain scores to the composite score is illustrated in Figures 11 and 12.

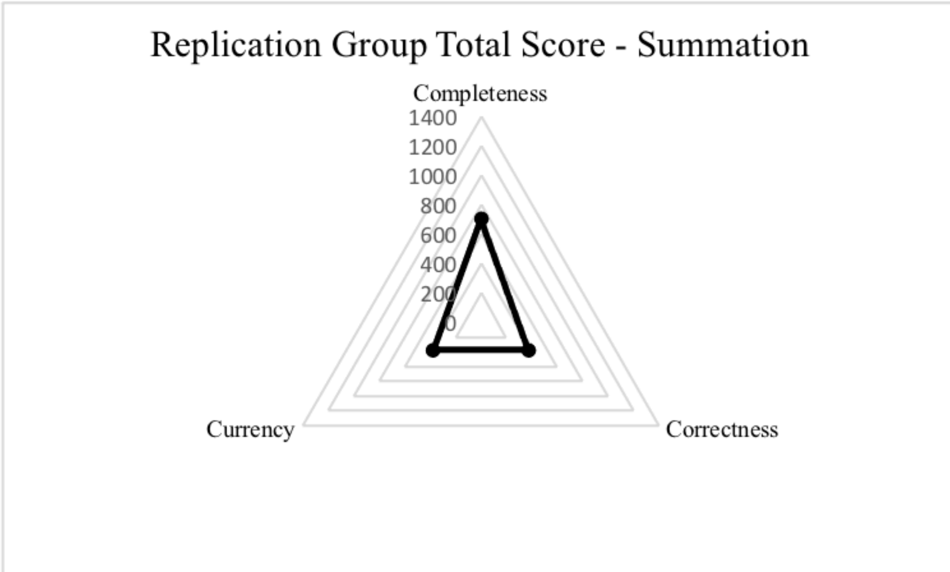

Figure 11: Replication group total component scores - Summation method

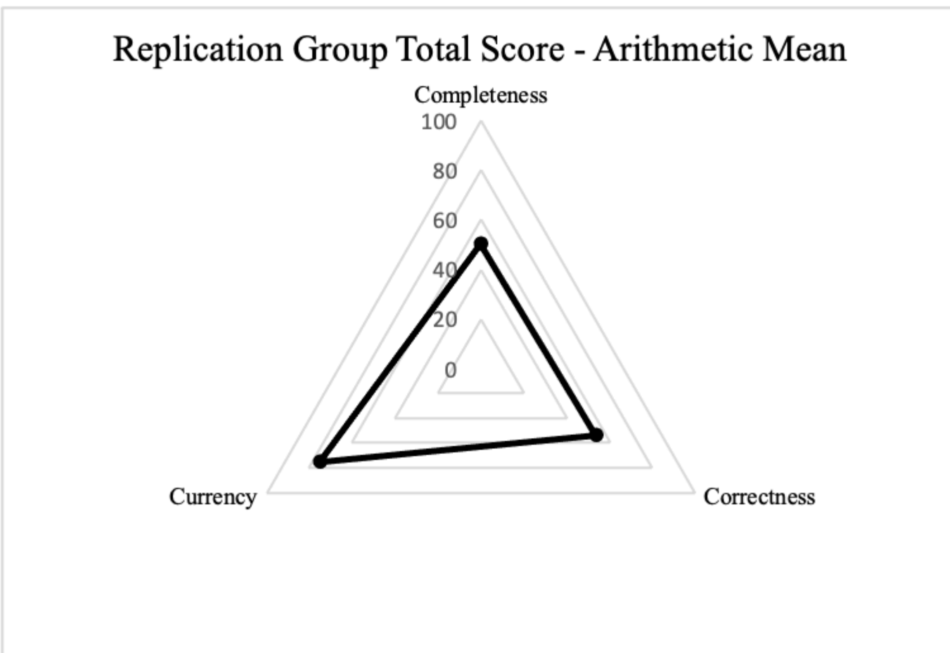

Figure 12: Replication group total component scores - Arithmetic mean method

4.5 Comparison of Development and Replication Groups

The scores from the Development and Replication Groups are summarized in Table 12 to facilitate comparison. A visual comparison of the relative arithmetic mean and summation results for the domain scores of the Development and Replication Groups can be found in Figures 13 and 14. The radar charts show that the scores on each domain are extremely similar between the Development and Replication Groups regardless of the composite score method, summation or arithmetic mean, used. The grey line represents the Development Group and the black line, representing the Replication Group, tracks directly on the grey line.

Table 12: Comparison of Development and Replication Groups

Domain	Method	Development Group	Replication group
Completeness	Summation	706.8	704.4
	Arithmetic Mean	50.5	50.3
Correctness	Summation	372.4	374.7
	Arithmetic Mean	53.2	53.1
Currency	Summation	386.4	375.7
	Arithmetic Mean	77.2	75.1
Total score	Summation	1465.5	1454.6
	Arithmetic Mean	60.3	59.6

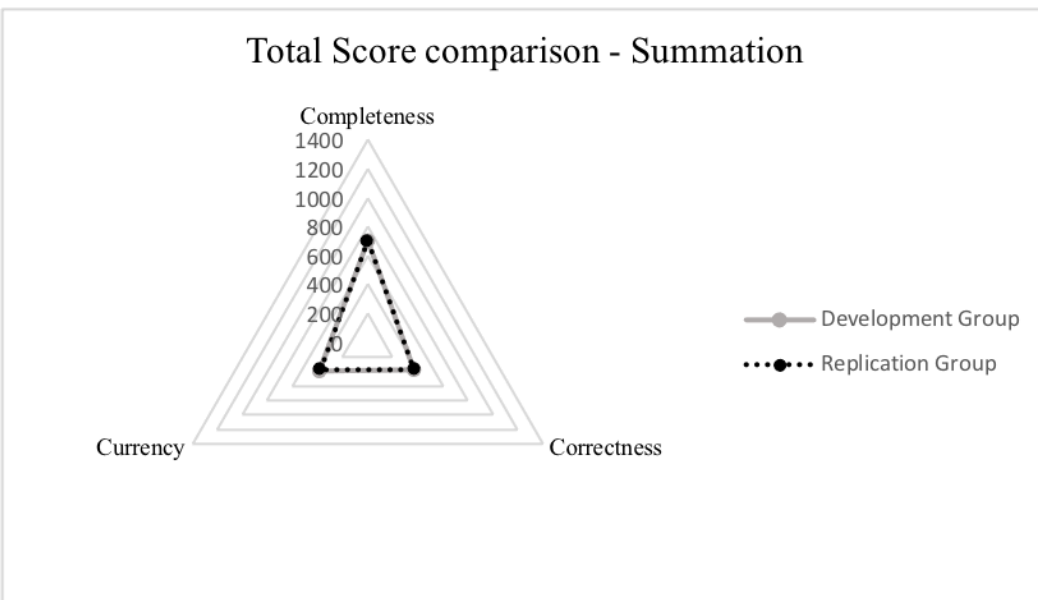

Figure 13: Comparison of Development and Replication groups using the summation method

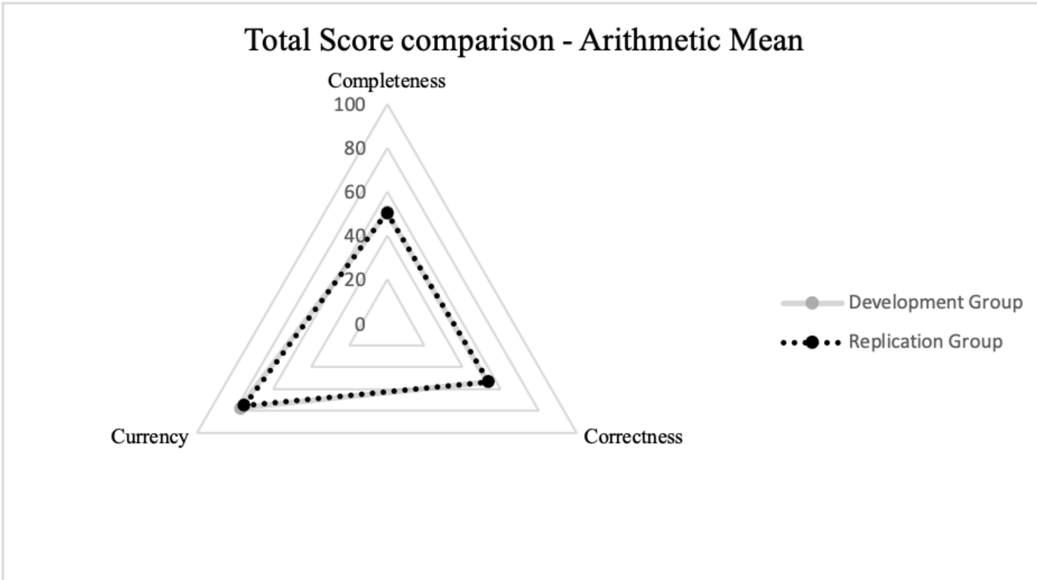

Figure 14: Comparison of Development and Replication groups using the arithmetic mean method

Chapter 5

5 Discussion

This thesis met the objectives of developing three domain-level scores and one composite score to describe data quality for primary care EMR datasets. Scores were developed at multiple assessment method levels (e.g. sensitivity values for hypertension) representing different measures (e.g. sensitivity) and combined to create domain-level scores (e.g. Completeness). These assessment method values allowed examination of variation between assessment methods within domains and provided domain-level scores (e.g. completeness) which can be compared across other EMR data sets. Finally, the domain level scores were combined to yield one single composite score that can be consistently calculated across other EMR data sets.

There were several challenges that were encountered in the development of these scores. These include missingness in the data, potential omission of data due to formatting conflicts between how the data were stored and the format needed for use in the analyses. These were limitations inherent to the CPCSSN data and which would be encountered in any use of EMR-derived data regardless of source. Given the novelty of this work, there is limited research on the interpretation of these types of data quality scores.

In the context of the current research landscape, primary care EMR data are a rich source of information for researchers but the scale and complexity can make assessment of the quality of these data overwhelming. The scores developed in this thesis, especially the composite score, will provide a roadmap for researchers who want to study EMR data quality assessment more robustly. As well, it is recommended that the calculation of these scores become standard metrics all researchers report when publishing research using EMR data sets. The use of this composite score would also map into section 19.1 of the “REporting of studies Conducted using Observational Routinely collected health Data” (RECORD) statement, which addresses discussions of the implications of using data to answer research questions for which they were not collected. RECORD is a supplement to the "Strengthening the Reporting of Observational Studies in Epidemiology" (STROBE) guidelines[57]. RECORD was developed to extend STROBE to address concepts specifically related to the secondary use of routinely collected health data, such as EMR data, in research[57].

5.1 Overview of results

5.1.1 Development of Domain Level and Composite Scores

In the score Development Group, there was a large degree of variation between the results of the assessment methods within domains, and between the domain scores. Domain scores were created so as to not obscure sources of variation, which might be more strongly driven by one domain, and to allow closer examination of the different aspects of data quality.

5.1.1.1 Completeness Domain

Four measures containing 14 assessment methods were used to represent the domain of completeness. In the domain of completeness, sensitivity scores were generally lower than in the other measures of completeness, for example, consistency of capture.

Sensitivity scores were lowest for the test condition of obesity at 4.4% This indicates that very few of the cases of obesity identified by applying the reference standard test condition definition were in patients who also had the code for obesity in their billing data. The reference standard test condition definition identified many more patients with obesity than the billing data did. This discrepancy cautions that using billing data to identify patients with obesity could be problematic; this has important implications for researchers who study obesity using secondary data sources. It is possible that this discrepancy could be caused because a patient's reason for visiting their doctor may not be obesity and so the billing code recorded would be for the reason for the patient's visit rather than the patient's obesity which, while co-occurring would most often be unrelated. This could also indicate that obesity billing codes are not a good target when examining data quality; definitions which use other fields, like the reference standard used here, may be preferable.

The highest sensitivity value was for asthma (34.5%) although it was still a relatively low score. In all but two conditions, hypothyroidism and urinary tract infection, using the reference standard test condition definitions identified more patients with the condition than the billing data. It is possible that there is something different about the way in which hypothyroidism and urinary tract infection are diagnosed, treated, or billed, which makes the billing data better for capturing cases. It is also possible that the test condition definitions do not perform as well for these conditions. Finally, it should be noted that, in comparison to the other test conditions, the sample sizes identified using both the test condition definition and the billing data for hypothyroidism and urinary tract infection were small. These very small sample sizes could be artificially biasing the results.

Completeness of EMR data has been assessed in different ways. A previous study by Singer et al. assessed completeness of primary care EMR data by examining billing code and problem list entries with different results than this thesis[27]. It was found that completeness scores were high for hypertension and diabetes but slightly lower for hypothyroidism and asthma[27]. These results differ somewhat to what was found in this thesis where the completeness score related to diabetes was lower and the completeness score related to asthma was highest. This difference in results could be due to actual differences in completeness of the EMR data as well as differences in recording practices, or methods of assessing completeness. This thesis used a test condition definition as the reference standard and calculated sensitivity whereas Singer et al. used billing codes as a reference standard, compared only to the problem list and did not calculate sensitivity values as their assessment method[27]. The use of billing codes as the reference standard may explain the lower scores because billing codes will not necessarily be recorded for every problem discussed during a visit to a care provider.

“Consistency of capture” scores were relatively high; “percentage of patients who visited in past year with 1 or more prescribed medications” had the highest score (92%). The lowest score was for “percentage of patients with 1 or more allergy record entries” (40.6%), although this is still relatively high. This lower value is likely due to a failure to enter “no allergies” in the patient record when patients have no allergies. Scores for “Recording of blood pressure, height, and weight” were high and generally similar to each other. Similarly, for the measure “Recording of blood pressure among patients requiring a blood pressure measurement” scores were very high, higher than for the more general measure “Recording of blood pressure, height, and weight”. This makes sense as, in patients to whom blood pressure control is critical, higher levels of recording than for the general population of patients would be expected, due to closer monitoring. Similar rates of recording of blood pressure, height, and weight were found in patients with diabetes and in patients with hypertension, both conditions in which blood pressure monitoring is important.

5.1.1.2 Correctness Domain

Two measures containing seven tests were used to represent the domain of correctness. For the positive predictive value measure, results were extremely variable. The highest score was for the test condition of hypertension (89.4%). This indicates that almost all cases of hypertension identified through billing code entries were also identified using the reference standard test condition definition. For patients with diabetes, the majority of cases identified using the billing code were also identified using the reference standard test condition definition (84.7%). Notably, while the sensitivity value for obesity was extremely low, the positive predictive value was much higher (41.7%). This higher value indicates that the patients who have an entry of obesity in their billing data are more likely to also be identified by the reference standard test condition definition than the other way around. There were far fewer patients identified by billing code for obesity than were identified through the reference standard test condition definition so, given that there is a degree of overlap, positive predictive values being higher than sensitivity values is expected. Extremely low positive predictive values for urinary tract infection (1.7%) indicates that almost none of the cases of urinary tract infection entered in the billing table were identified through the test condition definition. It is possible that, as previously stated, the small sample size may be impacting the results. Alternately, there may be something specific to the way this condition is recorded that is causing the discrepancy. “Unlikely combinations of age and specific procedures” was calculated as the percentage of patients over the age of 10 who received a tetanus toxoid conjugate vaccine subtracted from 100 and had a result of 93.2%. This high value indicates very few patients met the criteria for the assessment method. The original value was subtracted from 100 to allow all scores to have the same directionality; higher scores indicating better data quality. The score on this assessment method might be high because lower scores would be generated by the incorrect recording of age, by recording the wrong vaccination in a patient’s record, or by a patient actually receiving an uncommon vaccine for their age group.

5.1.1.3 Currency Domain

Five measures containing five tests were used to represent the domain of currency. In the domain of currency, for “Timeliness of weight recordings for patients with obesity”, values were extremely high (99.5%); this was the highest score calculated in the Development Group. This high score might have been impacted by the fact that one potential component of the reference standard test condition definition used to identify patients with obesity was the presence of a BMI over 30 in the patient's record. Having a BMI recorded would necessitate a weight recording being taken, although not necessarily within a year of their last visit. Within the currency domain, the score for "Timeliness of visit for antenatal care" was high. Scores for "Timeliness of blood pressure, height, and weight recordings" were consistently relatively high. The highest score was on "percentage of adult patients with one or more blood pressure values recorded less than one year prior to their last visit in the database" (67.6%) and the lowest score was on "percentage of patients with one or more height values recorded less than one year prior to their last visit in the database" (59.4%), although this is still relatively high. It might be anticipated that adult patients are more likely to have their blood pressure recorded recently in their record then height as height would be expected to change relatively little in adulthood.

5.1.1.4 Comparison of Domain Scores

The domain scores within the Development Group can only be accurately compared to each other using the scores calculated with the arithmetic mean method as the differing numbers of assessment methods impacts the scores when the summation method is used. As previously discussed the summation scores bias the scores towards the domains with more assessment methods because equal weight is given to each assessment method.

The domain scores calculated using the arithmetic mean method are varied. The highest domain score was currency (77.2). Both the domain of completeness (50.5) and the domain of correctness (53.2) had much lower scores. The lower score on completeness is driven by the generally low sensitivity values while the lower completeness score is significantly impacted by the positive predictive value for urinary tract infection, which was close to zero.

5.1.2 Comparison of Development and Replication Groups

The Development Group and the Replication Group were extremely similar. While there was some variation between scores on the individual assessment methods, the domain scores were very similar across groups using both the summation and arithmetic mean methods. The overall composite scores were also similar across groups for both the summation method (Development: 1465.5, Replication: 1454.6) and for the arithmetic mean method (Development: 60.3, Replication: 59.6). This consistency across the two similar groups indicates good reliability in the scoring system. Data splitting was used as a method of determining reliability. No external validation, that is, testing the scoring system using another data source, was used in this thesis. A further project could be undertaken to examine the scoring system's reliability in this context. It was not possible to validate the scoring system by comparing the CPCSSN data against the original patient EMR charts because that would require manual chart review, which was not possible because of the inability to obtain identifiable EMR records from the participating physicians and networks in CPCSSN. Further, this manual review is outside the scope of the master's thesis.

5.1.3 Comparison to Terry et. al

Given that the assessment methods, measures, and domains used in the current project were based on those used by Terry et. al[8], a comparison to their results is instructive. The sample size for this thesis was larger than that used by Terry et al. (919,551 patients versus 47,868 patients). This thesis used national data whereas Terry et al. used regional data from south western Ontario. The demographics of the patients included in the dataset used in this thesis were very similar to those reported in Terry el al.[8].

Based on the results of Terry et. al[8], some of the current findings were expected, but some deviated significantly from that found by Terry et al. Within the domain of completeness, the sensitivity values calculated in this thesis were generally lower than the values calculated in Terry et. al[8]. The most significant differences were in the scores calculated for the assessment method “sensitivity values for diabetes mellitus” which were generally quite high in Terry et. al[8] but were much lower in the present thesis and in the scores for the assessment method “sensitivity value for urinary tract infection, which was also much higher in Terry et. al[8]. The “sensitivity values for diabetes mellitus” have a fairly large variance within the study by Terry et al.[8] so it is possible that this particular assessment method shows large variance by dataset. The scores in consistency of capture measure within the domain of completeness were very similar to the values calculated in Terry et. al[8]. The largest difference in the consistency of capture measure was on the assessment method “percentage of patients who visited in past year with 1 or more prescribed medications”. These scores were higher in the present thesis in general, though within Terry et al. there is a lot of variance and the score calculated from dataset C is closer to the scores in this thesis than to the scores from the other two datasets in Terry et al[8], possibly indicating that this assessment method varies significantly by dataset. The measures “Recording of blood pressure, height, and weight” and “Recording of blood pressure among patients requiring a blood pressure measurement” had very similar scores in both this thesis and in Terry et. al[8].

Within the domain of correctness, some of the positive predictive values were similar between this thesis and Terry et. al and some differed significantly. The score calculated for the assessment method “positive predictive value for urinary tract infection” was similarly low in both the present thesis and Terry et. al[8]. The positive predictive values calculated for diabetes mellitus, hypertension, and asthma were all higher in the present thesis compared to Terry et. al, with the score for the positive predictive value for asthma in particular being much higher[8]. This discrepancy may be due to some issue in the data used in Terry et al. as one of the datasets did not have information on asthma and so no score was calculated and one of the datasets had no patients who both met the reference standard for asthma and had an asthma billing code in their record which was unusual. On the “positive predictive value for obesity” assessment method, the scores calculated for the three datasets in Terry et al. are extremely varied (83.5, 4, and 61.8)[8] which makes it difficult to compare the scores calculated in the present thesis. On the measure “unlikely combinations of age and specific procedure”, the values calculated in the present thesis were slightly higher than those in Terry et. al (6.8% and 4.2% compared to 0%)[8]. This is comparing the percentage of patients over 10 year of age with a tetanus toxoid conjugate vaccine rather than subtracting this number from 100 which is what was used in the score development. In this case it is possible that, with the much larger dataset used in this thesis, there was more opportunity for errors in recording, such as mistakenly recording the date a record was transferred to the system as the date a vaccination was given, or that more edge cases appeared as this vaccination is occasionally given to patients over 18 who were not given routine vaccinations as a child.

Within the domain of currency there is some variation between the scores calculated in the present thesis and those of Terry et. al. On the measure “timeliness of weight recordings for patients with obesity” the scores calculated in the present thesis are higher than those found by Terry et. al[8]. Likewise, on the measure “timeliness of visit for antenatal care” the score calculated in the present thesis were much higher than those in Terry et. Al, although the scores within Terry et al. were varied as well so this may be a metric which naturally shows a lot of variance[8]. On the measure "Timeliness of blood pressure, height, and weight recordings" the values in the present thesis are similar to those in Terry et. al[8] although the scores on the assessment method "percentage of patients with one or more height values recorded less than one year prior to their last visit in the database" were slightly higher in the present thesis.

5.2 Comparison to Other Literature Regarding Data Quality Assessment Approaches

This thesis expands on previous literature, by combining multiple assessment methods to create domain scores, by looking a multiple domains of data quality, and by combining these domain scores into a composite score that reflects data quality. Other EMR data quality projects have focused on a specific domain of data quality, such as completeness and examined this metric individually for various health conditions[6].

Other data quality literature has focused on assessing concordance between EMR data and external disease registries[12] or administrative datasets[7]. While it can still be important for researchers to examine comparability by examining EMR data in relation to other data sources, e.g. difference in recorded disease rates within the same population by database[7], a relevant reference standard must be accessible to do so. As previously discussed, the data quality scores generated in this thesis fall under the verification data quality assessment context as described by Kahn et al., meaning that no external comparators are required to generate the data quality scores[22]. Data quality frameworks that relate to the validation assessment context, evaluating data quality using external comparators, could be used as a complement to the score developed in this thesis.

Data quality assessment methods reported in the literature are frequently conceptual[10,21,22], rather than generating ratings or numeric scores. They often are not specifically for use by researchers and require expertise and knowledge which may make common adoption by researchers from unlikely. The creation of the scoring system in this thesis considered both the creation of concrete, rather than theoretical outcomes, and ease of use by researchers.

Overall, data quality assessments are frequently extremely specific to the data being assessed, conceptual rather than score driven, or require an outside comparator. The domain and composite scores developed in this thesis exclude the need for an outside comparison standard to increase usability and generalizability to all primary care EMR datasets. By combining multiple domains and multiple assessment methods within the domains, the composite score gives a multidimensional measure of data quality.

5.3 Challenges and Limitations

5.3.1 Missingness

The data from the CPCSSN dataset, used in this thesis analysis, have a degree of missingness, as is to be expected in data not collected for research purposes. In the examination of the patient demographics, it was found that 0.2% of patients were missing sex, 0.8% of patients were missing a birth year, and 10% of patients were missing a location code (denoting whether the patient resided in an urban or rural area). The missingness of sex and age were very low but the missingness in location code was substantially higher. However, the location code variable was not one required in any of the analyses. It is possible that, because this variable is not a routine part of patient care, there is not as much focus on entering it in an electronic medical record.

There were several steps undertaken in preparing the data for analysis, which could have led to the omission of data that is actually in the medical record, potentially leading to misleading results. The first is the use of free text field searches; these fields were used as part of the reference standard to identify patients with the medical conditions of interest or to identify patients who had a specific lab test or vaccination. As implied in its name, free text has no standardized formatting and can include multiple names of the same entity or spelling errors. Due to the size of the dataset, it was not feasible to go through each entry in the tables; as an example, the lab table alone contained 151,975,707 entries. Significant efforts were made to identify multiple permutations referring to the same condition or test. One example was, when searching for patients who had had a lab test for thyroid stimulating hormone levels, the free text was searched for “Thyroid-Stimulating Hormone”, “Thyroid Stimulating Hormone”, “Thyroid Stimulating Hormone (TSH)”, “Thyroid-Stimulating Hormone (TSH)”, and “TSH”. However, it is possible that there were permutations that were missed including potential spelling errors. These patients would not be included in the reference standard and would therefore decrease the calculated sensitivity and positive predictive values, thereby decreasing the scores artificially.

The second potential omission of data from the analyses comes from the use of the “TRY_CONVERT” command in the analysis code. Several of the reference standards used in the analyses required identifying patients who had specific lab results (e.g. patients with a fasting blood glucose result of 7.0 or greater). Within the SQL server, the numeric lab results in the “Test Result” column were stored in the “nvarchar” format. The format “nvarchar” indicates variable-length Unicode string data. To simplify, this means the “nvarchar” format can accommodate letters, numbers, and other characters including punctuation and other special characters. For the purposes of the analyses, the data in the lab results column had to be analysed as numbers, so the format had to be changed from a character string format to a numeric format. Because the “nvarchar” format has the potential to contain non-numeric characters, the “TRY_CONVERT” command had to be used instead of the “CONVERT” command. The trade-off of using “TRY_CONVERT” is that any results which could not be converted to numbers (e.g. if an error was made in the original data entry and the letter “o” was entered in place of the number “0”) would be excluded from the analyses. It was not possible to check the exact impact the excluded variables had on the results because in order to determine if a patient met the criteria of the operationalized measure, the lab result had to be examined and, in order to examine the results, the results had to be converted. The overall percentage of results from the lab table, containing non-numeric characters (which would be non-convertible) was 0.004% (570,093 records of 151,973,707) which is a negligible amount. To examine how this omitted data impacted the specific lab test results of interest, the lab table would have to be manually reviewed. Due to the size of the table, this was not feasible for the present thesis, which is a common issue when using data of this size. In research using large datasets, such as those that can be derived from EMRs, manual reviews of the data present challenges that make them impossible to conduct for practicality. It is therefore a good step in the early research process to examine whether the data can be manipulated in the manner required for the desired tests and procedures and how much data might be excluded from analysis based on incompatibility. In the case of this thesis, given the extremely small number of non-convertible records, it is unlikely there was a major impact on the score creation.

5.3.2 EMR Structure

The data used in the creation of the domain and composite scores was provided by CPCSSN. CPCSSN data are extracted from multiple EMR software packages[38] and CPCSSN uses algorithms to clean and validate the collected EMR data[39]. These two factors are important to note as they were related to the development of the current scores and they will be important considerations in the score's future use. Because the score was developed and tested on already manipulated data, variability in the score may have been produced from variability at the health care provider data entry level or from variability introduced by CPCSSN processes at either the network or overall CPCSSN level. CPCSSN data processing includes cleaning the data (e.g. standardizing formats of dates, removing duplicates), standardizing units (e.g. conversion of inches to centimetres), mapping the data onto the CPCSSN structure (e.g. standardizing classification and coding systems, structuring the data as CPCSSN tables), and the creation of some new variables from the raw data (e.g. case identification). This extensive data processing increases ease of use and allows integration of data from multiple EMR systems into a single cohesive dataset. However, it introduces the possibility that some variation in the data is not inherent to the raw data but is created by how well a given EMR maps onto the CPCSSN structure. It is possible that some EMRs are structured in a way that better aligns with how CPCSSN structures the data. These EMRs would potentially provide better quality data within CPCSSN than EMRs that were more dissimilar in structure and therefore required more manipulation and more potential opportunities to lose data in cleaning. Further, some EMRs will simply be structured in a way that provides better quality data due to how their input is structured (e.g. open text fields versus drop down lists, automatic prompts). In future use of this score, it will be important to investigate whether there is a difference in the performance of the score dependent on how the data were manipulated prior to the score being applied. Because in data cleaning some records are discarded or altered to align with the needs of researchers, it would be expected that heavily processed data, such as CPCSSN data, would score more highly than raw data, especially on the domain of completeness. No EMR data can be used for research without some level of processing; it is essential that researchers carefully document what is done to the data so that their research is replicable and transparent given that manipulation of the data has the potential to fundamentally change relationships within the data.

5.3.3 Unentered Data

In the calculation of the score, and in any manipulation of data, it is only possible to work with the data that is present. Not all patient data that is collected or that would be discussed in the duration of a patient visit will be recorded in the patient's EMR. There are many potential reasons for this including the structure of an EMR system, mandatory versus optional fields, ease of use, and provider discretion regarding what information is important or relevant. This factor of unentered data most obviously impacts the domain of completeness. The direct impact can be seen on assessment measures like "% of patients with 1 or more allergy record entries". In the allergy record example, it is likely that a relatively high percentage of the patients with no allergy record entry have no known allergies but that this was not necessarily seen as a relevant entry in the allergy record section of the EMR. Unentered data can also impact the other domains of data quality as it removes potentially important data points from analyses that would be used in the calculation of scores for those domains. Further, if there are patterns in the failure to enter data, such as if it is more common to fail to enter data for people with a specific health condition, this could systematically impact the score. Outside of the completeness measures, which partially speak to unentered data, it is difficult to measure how this issue impacts the score and how widespread it might be. Because the EMR system that physicians use might impact the amount of data that is or is not entered and also in what areas data might not be entered, further research could focus on the completeness domain score to examine the relationships between EMR system and unentered data.

5.4 Use of This Score

One of the main challenges in developing data scoring systems is that data quality is always relative, especially when the data are secondary and not developed for research purposes. Data quality, in terms of research, can be conceptualized as having two important aspects: the underlying quality of the data, which is what the score developed in this thesis attempted to represent, and the fitness of the data for purpose i.e., how well the data are able to speak to the research question of interest. The present scoring system aims to be useful to any researcher using primary care EMR data, irrespective of the particular research topic being investigated. By focusing on the domains of data quality – completeness, correctness, and currency – previously identified and explored by data quality researchers, the score attempts to put a numeric value on the absolute quality of the data being examined.

However, this score is not intended to be used as the only exploration of data quality by researchers using primary care EMR data. The score should be used in conjunction with examination of fitness for purpose. Fitness for purpose is a difficult area of data quality to standardize because it is specific to the individual research question. While it may be preferable to rely on a single score to indicate data quality, the fitness for purpose aspect should always be considered in data quality conversations.

5.4.1 Score Interpretation

The composite score is presented as a single numeric value or a series of numeric values if looking at the individual domain scores. It is difficult to interpret the scores, both the domain scores and the overall composite scores calculated in this thesis. In an absolute sense, using this scoring method, the highest achievable score using the summation method would be 2600 and the highest achievable score using the arithmetic mean method would be 100. In comparison to the potential perfect scores, the calculated scores of the Development Group (summation: 1465.5, arithmetic mean: 42.5) and Replication Group (summation: 1454.6, arithmetic mean: 43.2) are not particularly high. However, the real value of a scoring system like this one is in the ability to use it as a point of comparison. EMR data are not collected with research purposes and often have deficiencies that would not be expected in research data. EMR data will never be perfect in terms of data quality due to how they are created. Rather than solely judging the absolute score, it is more appropriate to compare various datasets, using a standard method like this scoring system. By comparing the data used for a particular research study to datasets used in other studies, a picture of where data is deficient can emerge. It is possible that a score of 42.5 is quite high in the space of EMR data used in research but this is not possible to know until there are many points of comparison. Because this score allows separation of the domains of data quality, it is possible to interpret the individual domain scores in comparison to each other when using the arithmetic mean method. In this thesis, the currency of the data was found to be quite high (75.5 and 73.4) compared to completeness (39.4, 40.1) and correctness (32.7, 34.3) which were lower than currency but similar to each other.

Having explored both the arithmetic mean and summation methods in this thesis, if this score were to be adopted widely, the arithmetic mean calculation for domain and composite scores is recommended. The summation method is straightforward to use, shows the relative numbers of assessment methods in each domain through the different possible maximum domain scores, and allows comparison of scores between datasets. However, the arithmetic mean method was deemed preferable because it allows comparison not only between datasets but also between domains within datasets due to the theoretical maximum score always being 100 regardless of the number of assessment methods used.

Were the present dataset to be used for future research, it would be prudent to consider that completeness and correctness are areas that could impact the research outcomes. The use of the domain scores in this way could alert researchers to areas of concern in their data and could prompt further investigation if one domain of the data is significantly different from others,

5.4.2 Role of Data Quality Scores in Primary Care Research

Primary care EMR data are becoming more and more available to Canadian researchers. Networks such as CPCSSN and its component networks provide easier access to these data than has previously been possible. This access has the potential to accelerate the growth of research using primary care EMR data and to introduce researchers who have not previously used datasets of this type and magnitude into this research space. While data access of this kind provides many exciting opportunities for research, which can improve public health, policy making, and patient's lives, it is important to remember the potential pitfalls of using data not collected for research purposes. However, while EMR data will never be perfect they are still a valuable resource for researchers interested in answering questions that require primary care clinical data. The desire for ideal data should not prevent researchers from using less than perfect data however, this does not mean data quality is irrelevant. The scoring system developed in this thesis aims to be a checkpoint for researchers, both to allow an easy entry point into data quality examination, an area that many researchers do not address, and to serve as a reminder of the necessity of such examination. The widespread use of this scoring system in the primary care EMR research space has the potential to allow researchers to compare relative data quality between projects. Further, examinations of drivers of data quality in EMR datasets could be assisted through using this standardized scoring method, leading to efforts to improve data quality. As more and more research takes place using data not derived for research purposes, it is more important than ever to examine the quality of the data and to remember the potential impacts of data quality on research outcomes.

5.5 Future Directions

There are many potential areas of expansion related to the present thesis. As previously discussed, the composite score has the most utility as a point of comparison between various EMR datasets and projects. If the score is widely adopted, projects could be undertaken to compare both the scores themselves as well as the underlying drivers of the scores. Further, the correlates of both the composite score and the domain scores could be explored, examining aspects such as what EMR system was used or how the data was processed. In this dataset, there were 11 different EMR systems used. There was large variation in how many providers used each EMR system with some being used by a quarter of providers and some being used by very few. EMR systems are not typically chosen by health care providers or healthcare centres based on the quality of data they would provide to researchers and variance in data quality by EMR system might be expected. The scoring system described here could be used with data similar to that found in this thesis to examine potential variability in data based on EMR system and what areas of the EMR drive this variability; for example, if one EMR system scores much lower on an assessment method or domain. This exploration could provide insight into why certain variations in data quality might exist and could be expanded to look at not only the scores but the structure of the EMR system and how data input and storage function contribute to variability. Finally, the accuracy of the score as a metric of data quality could be examined, possibly through a chart review project that involves comparing expert chart review scores with the scores produced by the current algorithm.

The score itself could also be expanded upon. The score as it currently exists uses an equal weighting scheme for each of the three examined domains. In the future, multiple weighting schemes could be developed which differentially prioritize the three domains. These weighting schemes could have utility for researchers with projects that emphasized one domain over others. For example, if completeness was of particular interest or importance over correctness and currency, a system of weights which gave more weight to completeness could be used. Such a system could be relatively simple; for example, giving the prioritized domain twice the weight given to each of the other two, or more complex, for example three differential values based on expert opinions of each domain's relative importance. The main drawback of introducing variable weighting schemes is that there would be a loss of comparability between projects which used different weighting schemes. Additionally, accurately determining the relative importance of each domain in relation to a specific project could be extremely complex and may not add much value to the score. However, if desired, researchers could calculate both the standard score and a weighted score that is specific to their project needs.

5.6 Conclusion

This thesis sought to create a unified scoring system for primary care EMR data that could be used by researchers to assess data quality. The created score was based on previous EMR data quality research and combined assessment methods of aspects of data quality falling under the domains of completeness, correctness, and currency. Two methods were used to combine the assessment method values into domain scores and the domain scores into a composite score: summation and averaging using the arithmetic mean. The arithmetic mean method is preferred for use going forward because it allows comparison between domains within a single project as well as comparisons across projects. The score was replicated using data splitting and was found to be reliable. Future use of this score could improve data quality reporting by providing a straightforward method to follow for researchers using primary care EMR data.

References

1. Hodge T, Giokas D. EMR, EHR, and PHR – Why all the confusion? | Canada Health Infoway. Canada Health Infoway. Published April 7, 2011. Accessed June 22, 2021. https://www.infoway-inforoute.ca/en/what-we-do/blog/digital-health-records/6852-emr-ehr-and-phr-why-all-the-confusion

2. Chang F, Gupta N. Progress in electronic medical record adoption in Canada. *Can Fam Physician*. 2015;61(12):1076-1084.

3. Upshur REG, Tracy S. Chronicity and complexity. *Can Fam Physician*. 2008;54(12):1655-1658.

4. Hasan S, Padman R. Analyzing the Effect of data quality on the accuracy of clinical decision support systems: A computer simulation approach. *AMIA Annu Symp Proc*. 2006;2006:324-328.

5. Ministry of Health and Long-Term Care. *Patients First: A proposal to strengthen patient-centred health care in Ontario*.; 2015:24. http://www.health.gov.on.ca/en/news/bulletin/2015/docs/discussion_paper_20151217.pdf.

6. Singer A, Kroeker AL, Yakubovich S, Duarte R, Dufault B, Katz A. Data quality in electronic medical records in Manitoba. *Can Fam Physician*. 2017;63(5):382-389.

7. Greiver M, Barnsley J, Glazier RH, Harvey BJ, Moineddin R. Measuring data reliability for preventive services in electronic medical records. *BMC Health Serv Res*. 2012;12:116. doi:10.1186/1472-6963-12-116

8. Terry AL, Stewart M, Cejic S, et al. A basic model for assessing primary health care electronic medical record data quality. *BMC Medical Informatics and Decision Making*. 2019;19(1):30. doi:10.1186/s12911-019-0740-0

9. 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):144-151. doi:10.1136/amiajnl-2011-000681

10. Faulconer ER, de Lusignan S. An eight-step method for assessing diagnostic data quality in practice: chronic obstructive pulmonary disease as an exemplar. *Inform Prim Care*. 2004;12(4):243-254.

11. Michael Bowen and Francis Lau. Defining and evaluating electronic medical record data quality within the Canadian context. *ElectronicHealthcare*. 2012;11(1):e5-e13.

12. Sollie A, Sijmons RH, Helsper C, Numans ME. Reusability of coded data in the primary care electronic medical record: A dynamic cohort study concerning cancer diagnoses. *Int J Med Inform*. 2017;99:45-52. doi:10.1016/j.ijmedinf.2016.08.004

13. Canadian Institute for Health Information. *How Canada compares: Results from the commonwealth fund's 2019 international health policy survey of primary care physicians — Accessible Report*. CIHI; 2020:78.

14. Rubinowicz A, Vedel I, Sanche S, et al. A Portrait of electronic medical record use in primary care across Canada. *HRO-ORS*. 2016;4(2). doi:10.13162/hro-ors.v4i2.2463

15. Raymond L, Paré G, Ortiz de Guinea A, et al. Improving performance in medical practices through the extended use of electronic medical record systems: a survey of Canadian family physicians. *BMC Med Inform Decis Mak*. 2015;15:27. doi:10.1186/s12911-015-0152-8

16. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. *Health Inf Sci Syst*. 2014;2. doi:10.1186/2047-2501-2-3

17. Keshavjee K, Williamson T, Martin K, et al. Getting to usable EMR data. *Can Fam Physician*. 2014;60(4):392.

18. De Coster C, Quan H, Finlayson A, et al. Identifying priorities in methodological research using ICD-9-CM and ICD-10 administrative data: report from an international consortium. *BMC Health Serv Res*. 2006;6:77. doi:10.1186/1472-6963-6-77

19. Birtwhistle R, Williamson T. Primary care electronic medical records: A new data source for research in Canada. *CMAJ*. 2015;187(4):239-240. doi:10.1503/cmaj.140473

20. Thompson M, Ramsey MH. Quality concepts and practices applied to sampling—an exploratory study. *Analyst*. 1995;120(2):261-270. doi:10.1039/AN9952000261

21. United Nations Statistical Commission. *Guidelines for the Template for a Generic National Quality Assurance Framework (NQAF)*. United Nations, Statistics Division; 2012. Accessed June 22, 2021. https://unstats.un.org/unsd/statcom/doc12/bg-nqaf.pdf

22. Kahn MG, Callahan TJ, Barnard J, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *EGEMS (Wash DC)*. 2016;4(1):1244. doi:10.13063/2327-9214.1244

23. Bian J, Lyu T, Loiacono A, et al. Assessing the practice of data quality evaluation in a national clinical data research network through a systematic scoping review in the era of real-world data. *Journal of the American Medical Informatics Association*. 2020;27(12):1999-2010. doi:10.1093/jamia/ocaa245

24. Liaw ST, Guo JGN, Ansari S, et al. Quality assessment of real-world data repositories across the data life cycle: A literature review. *Journal of the American Medical Informatics Association*. 2021;28(7):1591-1599. doi:10.1093/jamia/ocaa340

25. Hogan WR, Wagner MM. Accuracy of data in computer-based patient records. *J Am Med Inform Assoc*. 1997;4(5):342-355.

26. Cadarette SM, Wong L. An introduction to health care administrative data. *Can J Hosp Pharm*. 2015;68(3):232-237.

27. 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):1107-1112. doi:10.1093/jamia/ocw013

28. Thuraisingam S, Chondros P, Dowsey MM, et al. Assessing the suitability of general practice electronic health records for clinical prediction model development: A data quality assessment. *BMC Medical Informatics and Decision Making*. 2021;21(1):297. doi:10.1186/s12911-021-01669-6

29. Price M, Davies I, Rusk R, Lesperance M, Weber J. Applying STOPP guidelines in primary care through electronic medical record decision support: Randomized control trial highlighting the importance of data quality. *JMIR Med Inform*. 2017;5(2):e15. doi:10.2196/medinform.6226

30. Centre for Studies in Family Medicine. DELPHI: Deliver Primary Healthcare Information. Published 2019. Accessed June 22, 2021. https://www.schulich.uwo.ca/familymedicine/research/csfm//research/current_projects/delphi.html

31. Song MK, Lin FC, Ward SE, Fine JP. Composite variables. *Nurs Res*. 2013;62(1):45-49. doi:10.1097/NNR.0b013e3182741948

32. Gerstein HC, Ramasundarahettige C, Bangdiwala SI. Creating composite indices from continuous variables for research: The geometric mean. *Diabetes Care*. Published online February 25, 2021. doi:10.2337/dc20-2446

33. Bobko P, Roth PL, Buster MA. The usefulness of unit weights in creating composite scores: A literature review, application to content validity, and meta-analysis. *Organizational Research Methods*. 2007;10(4):689-709. doi:10.1177/1094428106294734

34. McNeish D, Wolf MG. Thinking twice about sum scores. *Behav Res*. 2020;52(6):2287-2305. doi:10.3758/s13428-020-01398-0

35. Saary MJ. Radar plots: A useful way for presenting multivariate health care data. *Journal of Clinical Epidemiology*. 2008;61(4):311-317. doi:10.1016/j.jclinepi.2007.04.021

36. Feldman R. Filled radar charts should not be used to compare social indicators. *Soc Indic Res*. 2013;111(3):709-712. doi:10.1007/s11205-012-0028-6

37. Queenan JA, Williamson T, Khan S, 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-E32. doi:10.9778/cmajo.20140128

38. Canadian Primary Care Sentinel Surveillance Network (CPCSSN). Canadian Primary Care Sentinel Surveillance Network (CPCSSN). Published 2021. Accessed January 26, 2021. http://cpcssn.ca/

39. Canadian Primary Care Sentinel Surveillance Network. About Us – Canadian Primary Care Sentinel Surveillance Network (CPCSSN). Published 2021. Accessed January 26, 2021. https://cpcssn.ca/about-us/

40. 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). *International Journal of Epidemiology*. 2017;46(4):1091-1092f. doi:10.1093/ije/dyw248

41. Kuhn M, Johnson K. *Feature Engineering and Selection: A Practical Approach for Predictive Models*. CRC Press; 2019.

42. Steyerberg EW. Validation in prediction research: The waste by data splitting. *Journal of Clinical Epidemiology*. 2018;103:131-133. doi:10.1016/j.jclinepi.2018.07.010

43. CPCSSN-ERD-v4.0.4.pdf. Accessed January 27, 2022. http://cpcssn.ca/wp-content/uploads/2019/05/CPCSSN-ERD-v4.0.4.pdf

44. Sullivan GM, Feinn R. Using effect size—or Why the P Value is not enough. *J Grad Med Educ*. 2012;4(3):279-282. doi:10.4300/JGME-D-12-00156.1

45. Davis WS. Pseudocode. In: *The Information System Consultant's Handbook*. CRC Press; 1999.

46. Accuro EMR | Canada's number one single platform EMR. Accuro EMR. Accessed February 6, 2022. https://accuroemr.com/

47. Healthquest EMR - Simply Vital to the Health of Your Practice. Healthquest. Accessed February 6, 2022. https://www.healthquest.ca/

48. Collaborative Health Record | TELUS Health. TELUS. Accessed February 6, 2022. https://www.telus.com/en/health/health-professionals/clinics/collaborative-health-record

49. Home | Intrahealth. Accessed February 6, 2022. https://www.intrahealth.com/

50. Med Access EMR | TELUS Health. TELUS. Accessed February 6, 2022. https://www.telus.com/en/health/health-professionals/clinics/med-access

51. Medesync EMR | TELUS Health. TELUS. Accessed February 6, 2022. https://www.telus.com/en/health/health-professionals/clinics/medesync

52. OSCAR. Accessed February 6, 2022. https://oscar-emr.com/

53. Electronic Medical Records EMR | P & P Data Systems Inc. Accessed February 6, 2022. https://www.p-pdata.com/electronic-medical-record-emr/

54. PS Suite EMR | TELUS Health. TELUS. Accessed February 6, 2022. https://www.telus.com/en/health/health-professionals/clinics/ps-suite

55. EMR Overview - Purkinje. Accessed February 6, 2022. https://www.purkinje.com/en/solutions-for-clinics/emr-overview

56. Wolf EMR | TELUS Health. TELUS. Accessed February 6, 2022. https://www.telus.com/en/health/health-professionals/clinics/wolf

57. Benchimol EI, Smeeth L, Guttmann A, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. *PLoS Med*. 2015;12(10):e1001885. doi:10.1371/journal.pmed.1001885

Appendices

Appendix A: Summary of reviewed studies using CPCSSN data

Author	Year	Study Title	Study Objectives	Tables/variables used
Singer et al.[1]	2021	Who is asked about alcohol consumption? A retrospective cohort study using a national repository of Electronic Medical Records	Describe documentation of alcohol use recorded in EMRs and model factors associated with documentation	Patient sex, patient age, patient postal code, provider demographics, provider type, health conditions table, risk factor table
Barber et al.[2]	2021	Patients eligible and referred for bariatric surgery in southeastern Ontario: Retrospective cohort study	Evaluate the proportion of eligible individuals, within one health region in Ontario, who were referred for publicly funded medical and surgical weight-loss interventions	Patient sex, patient age, patient postal code, health conditions table, billing, encounter diagnosis, medications, lab table, exam (height, weight, bp)
Singer et al.[3]	2020	Prevalence of Physician-Reported Food Allergy in Canadian Children	Determine the prevalence of physician-reported food allergy in children using electronic medical record data from the Canadian Primary Care Sentinel Surveillance Network (CPSSN).	Patient sex, patient age, patient postal code, provider demographics, health conditions table, billing, encounter diagnosis, medications, allergy table, exam (height, weight, bp)
Black et al.[4]	2020	Development and evaluation of an osteoarthritis risk model for integration into primary care health information technology	Develop and evaluate a prognostic prediction model that estimates osteoarthritis risk based on primary care data.	Patient sex, patient age, health conditions table, risk factor table, billing, encounter diagnosis, medications, exam (height, weight, bp)
Ross et al.[5]	2020	Pelvic floor disorders in women who consult primary care clinics: development and validation of case definitions using primary care electronic medical records	Develop and validate symptom-based case definitions for urinary incontinence, fecal incontinence and pelvic organ prolapse in women, for use in primary care epidemiologic or clinical research.	Patient sex, patient age, health conditions table, billing, encounter diagnosis, medications, exam (height, weight, bp)
Puzhko et al.[6]	2020	Evaluating prevalence of patterns of prescribing medications for depression for patients with obesity using large primary care data (Canadian Primary Care Sentinel Surveillance Network).	Describe the prevalence and patterns of AD prescribing for patients with depression and comorbid obesity compared with normal weight patients, and to examine the association of prescribing prevalence with obesity class	Patient sex, patient age, patient postal code, health conditions table, billing, encounter diagnosis, medications, exam (height, weight, bp) , network id
Miyagishima et al.[7]	2020	Validation of care definition for speech and language disorders: In community-	Validate a case definition for speech and language disorders in community-dwelling older	Patient sex, patient age, health conditions table, billing,

Author	Year	Study Title	Study Objectives	Tables/variables used
		dwelling older adults in Alberta	adults and to determine the prevalence of speech and language disorders in a primary care population.	encounter diagnosis, exam (height, weight, bp)
Marrie et al.[8]	2020	Management of diabetes and hypertension in people with multiple sclerosis	Determine the adequacy of disease control and intensity of treatment for diabetes and hypertension in individuals with multiple sclerosis (MS) as compared to individuals without MS using the Canadian Primary Care Sentinel Surveillance System.	Patient sex, patient age, patient postal code, health conditions table, billing, encounter diagnosis, medications, lab table, exam (height, weight, bp)
Cave et al.[9]	2020	Validation a case definition for adult asthma in primary care electronic medical records	Develop and validate a case definition to identify adults with asthma who consult family physicians and to estimate the prevalence of asthma in that setting in Canada	Patient sex, patient age, health conditions table, billing, encounter diagnosis, medications, exam (height, weight, bp), network id
Bang et al.[10]	2020	Surveillance of concussion – related injuries using electronic medical records from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN): a proof-of-concept	Explore the feasibility of using data collated from CPCSSN for concussion surveillance purposes and examine trends over time.	Patient sex, patient age, patient postal code, health conditions table, billing, encounter diagnosis
Bello et al.[11]	2019	Prevalence and demographics of CKD in Canadian primary care practices: A cross-sectional study	Estimate the prevalence and range of severity of CKD in the Canadian primary care context; and (ii) describe geographic, sociodemographic, and clinical variations in CKD prevalence.	Patient sex, patient age, patient postal code, health conditions table, billing, encounter diagnosis, medications, lab table, exam (height, weight, bp)
Garies et al.[12]	2019	Prevalence of hypertension, treatment, and blood pressure targets in Canada associated with the 2017 American College of Cardiology and American Heart Association	Evaluate the potential change in the diagnosis, treatment, and control of hypertension in a Canadian cohort of patients with hypertension attending primary care practices using the ACC/AHA guidelines.	Patient sex, patient age, health conditions table, risk factor table, billing, encounter diagnosis, medications, lab table, exam (height, weight, bp) , medical procedures
Rigobon et al. [13]	2019	Impact of the Diabetes Canada Guideline Dissemination Strategy on the prescription of vascular protective medications: A retrospective cohort study, 2010–2015	Examine changes associated with the launch of the 2013 guidelines and additional dissemination efforts in the rates of vascular protective medications prescribed in primary care for older patients with diabetes and examine differences in the rates of prescriptions of vascular protective medications by	Patient sex, patient age, patient postal code, provider demographics, health conditions table, billing, encounter diagnosis, medications, lab table, exam (height, weight, bp), network id

Author	Year	Study Title	Study Objectives	Tables/variables used
			patient and provider characteristics.	
Kosowan et al.[14]	2019	Enhancing health surveillance: validation of a novel electronic medical record-based pediatric type 1 and type 2 Diabetes Mellitus case definition	Compose and validate an electronic medical records–based case definition for pediatric diabetes in primary care.	Patient sex, patient age, health conditions table, billing, encounter diagnosis, medications, lab table, exam (height, weight, bp)
O'Neill et al.[15]	2019	Agreement between primary care and hospital diagnosis of schizophrenia and bipolar disorder: A cross-sectional, observational study using record linkage	Identify a cohort of patients labeled as having schizophrenia and bipolar disorder, who had been seen in both settings (primary care and hospital), estimate labeling agreement, and determine patient factors associated with labeling agreement.	Patient sex, patient age, patient postal code, health conditions table, billing, encounter diagnosis, exam (height, weight, bp), network id
Perveen et al.[16]	2018	Metabolic syndrome and development of diabetes mellitus: predictive modeling based on machine learning techniques	Investigate: 1) the relationship between diabetes mellitus and individual risk factors of metabolic syndrome (MetS), in a non-conservative setting; 2) the prediction of future onset of diabetes using relevant risk factors of MetS; and 3) investigate the relative performance of machine learning methods when data sampling techniques are used to generate balanced training sets.	Patient sex, patient age, health conditions table, billing, encounter diagnosis, lab table
Phillips et al.[17]	2019	Assessing wellness in the wellchild check What about social and emotional development?	Determine whether Canadian children aged 4 to 6 received well-child checks; to explore the nature of these checkups in a large family practice; and to examine the merit of using parent questionnaires about child resilience as a means of introducing a discussion about social and emotional development into this checkup.	Patient sex, patient age, health conditions table, billing, encounter diagnosis, exam (height, weight, bp), network id, vaccine table
Greiver et al.[18]	2019	Trends in end digit preference for blood pressure and associations with cardiovascular outcomes in Canadian and UK primary care: a retrospective observational study	Study systematic errors in recording blood pressure as measured by end digit preference; determine associations between EDP, uptake of Automated Office BP machines and cardiovascular outcomes.	Patient sex, patient age, health conditions table, medications, exam (height, weight, bp) , site table

Author	Year	Study Title	Study Objectives	Tables/variables used
Abu-Ashour et al.[19]	2018	Diabetes and the occurrence of infection in primary care: A matched cohort study	Estimate the association between diabetes and infections occurring in primary care.	Patient sex, patient age, health conditions table, risk factor table, billing, encounter diagnosis, exam (height, weight, bp), network id
Drummond et al.[20]	2018	Developing and implementing linked electronic medical record and administrative data in primary care practice for diabetes in Alberta	Create an interactive diabetes dashboard using de-identified, linked, patient data from EMRs and administrative data sets.	Patient sex, patient age, patient postal code, health conditions table, risk factor table, billing, encounter diagnosis
Drummond et al.[21]	2018	Antidepressant and antipsychotic prescribing in primary care for people with dementia	Use data from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) to evaluate the prevalence of antidepressant and antipsychotic prescriptions among patients with no previous depression or psychosis diagnoses, and to identify the factors associated with the use of these drugs in this population.	Patient sex, patient age, patient postal code, provider demographics, health conditions table, encounter diagnosis, medications
Ehsani-Moghaddam et al.[22]	2018	Mucopolysaccharidosis type II detection by Naïve Bayes Classifier: An example of patient classification for a rare disease using electronic medical records from the Canadian Primary Care Sentinel Surveillance Network	Evaluate the Naïve Bayes Classifier algorithm as a simple method of identifying a “group” of patients with the highest likelihood of having MPS II as a rare disease with relatively common features using a “real” dataset.	Patient sex, patient age, health conditions table, billing, encounter diagnosis, network id
Bartlett et al.[23]	2017	Antidepressant prescription practices among primary health care providers for patients with Diabetes Mellitus	Describe the prescription of antidepressants for diabetic patients with a focus on medications suspected to impair glucose control.	Patient sex, patient age, health conditions table, encounter diagnosis, medications, lab table
Kalia et al.[24]	2018	Would you like to add a weight after this blood pressure, doctor? Discovery of potentially actionable associations between the provision of multiple screens in primary care	Discover and rank associations between the presence of screens to plan more efficient prompts in primary care.	Patient sex, patient age, health conditions table, risk factor table, billing, encounter diagnosis, lab table, network id, site table
Hurd et al. [25]	2018	Health and health service use of very elderly Newfoundlanders	Develop a better understanding of the current health status and health care use of the population of very elderly Newfoundlanders.	Patient sex, patient age, patient postal code, health conditions table, billing, encounter diagnosis, medications, lab table,

Author	Year	Study Title	Study Objectives	Tables/variables used
				exam (height, weight, bp), network id
Greiver and Kalia[18]	2018	Trends in systematic recording errors of blood pressure and association with outcomes in Canadian and UK primary care data: a retrospective observational	Study EDP trends, uptake of Automated Office BP (AOBP) measurement, and cardiovascular outcomes in the UK and Canada.	Patient sex, patient age, health conditions table, medications, exam (height, weight, bp) , site table
Aponte-Hao et al.[26]	2018	Developing a primary care EMR-based frailty definition using machine learning	Develop an operational definition of frailty using machine learning that can be applied to a primary care electronic medical record (EMR) database.	Patient sex, patient age, health conditions table, billing, encounter diagnosis, medications, lab table, referral table
Gagnon et al.[27]	2018	The impact of antidepressant therapy on glycemic control in Canadian primary care patients with diabetes mellitus.	Estimate the impact of Citalopram, Amitriptyline, Venlafaxine, Trazodone, and Escitalopram on glycemic control in Canadian primary care patients with diabetes.	Patient sex, patient age, health conditions table, billing, encounter diagnosis, medications
Mangin et al.[28]	2018	Legacy drug-prescribing patterns in primary care	Evaluated the proportion of legacy prescribing within antidepressants, bisphosphonates, and proton pump inhibitors.	Patient age, patient postal code, provider demographics, medications, network id
McAlister et al.[29]	2018	Use of direct oral anticoagulants in Canadian primary care practice 2010-2015: A cohort study from the Canadian primary care sentinel surveillance network	Studied DOAC prescribing patterns in Canada among patients managed in the outpatient setting by primary care providers and explored whether patient or provider factors were associated with the prescribed dose.	Patient sex, patient age, patient postal code, provider demographics, provider type, health conditions table, billing, encounter diagnosis, medications, lab table, exam (height, weight, bp) , site table
Weaver et al.[30]	2018	Association rule mining to identify potential under-coding of conditions in the problem list in primary care electronic medical records	Use association rule mining to develop association rules between hypertension, diabetes, and depression and other clinical information available in the EMR, such as other diagnoses in the problem list, billing codes, medications, and laboratory results.	Patient sex, patient age, patient postal code, health conditions table, billing, encounter diagnosis, medications, lab table, site table
Perveen et al.[31]	2018	A systematic machine learning based approach for the diagnosis of non-alcoholic fatty liver disease risk and progression	Develop machine learning based method in order to identify individuals at an increased risk of developing NAFLD using risk factors of ATP III clinical criteria updated in 2005 for Metabolic Syndrome (MetS). Validate the relative ability of	Patient sex, patient age, health conditions table, lab table, exam (height, weight, bp)

Author	Year	Study Title	Study Objectives	Tables/variables used
			quantitative score defined by Italian Association for the Study of the Liver (IASF) and guideline explicitly defined for the Canadian population based on triglyceride thresholds to predict NAFLD risk.	
Katz et al.[32]	2018	Identification of frailty using EMR and admin data: a complex issue	Develop an administrative data definition of frailty to facilitate clinical and health system planning and validate the definition by linking the administrative data to electronic medical records (EMR) data.	Patient sex, patient age, health conditions table, billing, encounter diagnosis, medications, lab table, exam (height, weight, bp)
Queenan et al.[33]	2017	The prevalence and risk for herpes zoster infection in adult patients with diabetes mellitus in the Canadian Primary Care Sentinel Surveillance Network	Develop and validate a case definition of Herpes zoster (HZ) diagnosis based on electronic medical records; determine a prevalence estimate for HZ in adult patients in the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) and assess the association between HZ and diabetes.	Patient sex, patient age, health conditions table, billing, encounter diagnosis, medications, lab table
Reyes et al.[34]	2018	Team-based comanagement of diabetes in rural primary care.	Explore clinical indicators among patients with diabetes in southern Alberta and assess changes over time, and to compare patients with diabetes attending a reference clinic, which had adapted its service model to address the specific needs of the patient population, with patients with diabetes attending comparison clinics in the same region.	Patient sex, patient age, health conditions table, lab table, exam (height, weight, bp) , network id
Greiver et al.[35]	2018	Agreement between hospital and primary care on diagnostic labeling for COPD and heart failure in Toronto, Canada: a cross- sectional observational study	Generate cohorts of patients seen in both settings labeled as having heart failure (HF) or COPD, to estimate agreement on recorded labels of COPD or HF, to determine patient factors associated with agreement on labeling, and to estimate the size of the populations of interest.	Patient sex, patient age, patient postal code, health conditions table, billing, encounter diagnosis, exam (height, weight, bp) , network id
Singer et al.[36]	2017	Informing antimicrobial stewardship: factors associated with inappropriate antimicrobial prescribing in primary care	Determine the frequency of potentially inappropriate antimicrobial prescribing in primary care practices in Manitoba, Canada and to assess the association between	Patient sex, patient age, patient postal code, provider demographics, health conditions table,

Author	Year	Study Title	Study Objectives	Tables/variables used
			potentially inappropriate antimicrobial prescribing and patient, prescriber and practice-related factors.	billing, network id, site table
Singer et al.[37]	2018	Prescribing and testing by primary care providers to assess adherence to the Choosing Wisely Canada recommendations: a retrospective cohort study.	Assess adherence to 4 Choosing Wisely Canada recommendations	Patient sex, patient age, provider demographics, health conditions table, billing, encounter diagnosis, medications, lab table, network id, site table
Bello et al.[38]	2017	A national surveillance project on chronic kidney disease management in Canadian primary care: a study protocol	Improve the understanding of how patients with chronic kidney disease (CKD) are managed in primary care and the variation across practices and provinces and territories to drive improvements in care delivery.	Patient sex, patient age, patient postal code, health conditions table, billing, encounter diagnosis, medications, lab table, exam (height, weight, bp) , medical procedures
Wong et al.[39]	2018	Can linked electronic medical record and administrative data help us identify those living with frailty?	Construct algorithms that can identify frailty using electronic medical record (EMR) and administrative data. Describe sociodemographic characteristics, risk factors, prescribed medications, use and costs of healthcare for those identified as frail.	Patient sex, patient age, health conditions table, billing, encounter diagnosis, medications, lab table, exam (height, weight, bp)
Abrams et al.[40]	2017	Adherence with epinephrine autoinjector prescriptions in primary care	Estimate primary adherence for epinephrine autoinjector (EA) prescriptions in primary care practices in Manitoba, Canada.	Patient sex, patient age, patient postal code, health conditions table, billing, encounter diagnosis, medications, network id
O'Brien et al.[41]	2017	Piloting electronic screening forms in primary care : findings from a mixed methods study to identify patients eligible for low dose CT lung cancer screening	Compare the acceptability and feasibility of using brief electronic versus paper screening forms to identify eligible patients at high risk of developing lung cancer in primary care.	Patient age, exam (height, weight, bp)
Brown et al.[42]	2017	Statin-prescribing trends for primary and secondary prevention of cardiovascular disease	Determine the proportion of patients receiving statins for primary or secondary prevention of cardiovascular disease (CVD), as well as to describe lipid-screening trends.	Patient sex, patient age, health conditions table, billing, encounter diagnosis, medications, network id
Coons et al.[43]	2017	Is glycemia control in Canadians with diabetes individualized? A cross-sectional observational study	Examine whether glycemic control varied by age and comorbidities in Canadian primary care.	Patient sex, patient age, health conditions table, medications, lab

Author	Year	Study Title	Study Objectives	Tables/variables used
				table, exam (height, weight, bp)
Loo et al.[44]	2017	Association between neighbourhood walkability and metabolic risk factors influenced by physical activity: A cross ;/sectional study of adults in Toronto, Canada.	Determine whether neighbourhood walkability is associated with clinical measures of obesity, hypertension, diabetes and dyslipidaemia in an urban adult population.	Patient sex, patient age, patient postal code, health conditions table, risk factor table, billing, encounter diagnosis, medications, lab table, exam (height, weight, bp)
Lix et al.[45]	2017	Chronic disease case definitions for electronic medical records: a Canadian validation study	Test the accuracy of chronic disease case definitions in EMR data from one CPCSSN site.	Patient sex, patient age, health conditions table, billing, encounter diagnosis, medications, lab table, exam (height, weight, bp)
Lix et al.[46]	2017	Automated Classification of Alcohol Use by Text Mining of Electronic Medical Records	Develop and validate an automated system to extract and classify patient alcohol use based on unstructured(i.e., free) text in primary care electronic medical records	Exam (height, weight, bp)
Morkem et al.[47]	2017	Recent trends in the prescribing of ADHD medications in Canadian primary care	Describe the prevalence and incidence of ADHD medication prescribing, by age and gender, from 2005 to 2015 in Canadian primary care.	Patient sex, patient age, medications
Morkem et al.[48]	2017	Trends in antidepressant prescribing to children and adolescents in Canadian primary care: a time-series analysis	Describe the trends and patterns of antidepressant (AD) prescribing to children and adolescents in Canadian primary care before and after the black-box warning in 2004.	Patient sex, patient age, medications
Oake et al.[49]	2017	Using electronic medical record to identify patients with dyslipidemia in primary care settings: international classification of disease code matters from one region to a national database	Assess the validity of the International Classification of Disease (ICD) codes for identifying patients with dyslipidemia in electronic medical record (EMR) data.	Patient sex, patient age, health conditions table, risk factor table, billing, encounter diagnosis, medications, lab table, exam (height, weight, bp) , network id
Aref-Eshghi et al.[50]	2017	Identification of dyslipidemic patients attending primary care clinics using electronic medical record (EMR) data from the Canadian primary care sentinel surveillance network (CPCSSN) database	Define the optimal algorithm to identify patients with dyslipidemia using electronic medical records.	Patient sex, patient age, health conditions table, billing, encounter diagnosis, medications, lab table
Singer et al.[51]	2017	Data quality in electronic medical records in Manitoba: Do problem lists reflect	Determine if the problem list (health conditions) in primary care electronic medical records (EMRs) accurately reflects the	Health conditions table, billing, medications, site table

Author	Year	Study Title	Study Objectives	Tables/variables used
		chronic disease as defined by prescriptions ?	conditions for which chronic medications are prescribed in the EMR.	
Ryan et al.[52]	2017	Methods to describe referral patterns in a Canadian primary care electronic medical record database: modelling multilevel count data	Establish methodology for determining referral rates from FPs to medical specialists using the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) EMR database.	Patient sex, patient age, provider demographics, health conditions table, site table, referral table, disease case
Birtwhistle et al.[53]	2017	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.	Link electronic medical record (EMR) data from Ontario patients in the CPCSSN with administrative data to assess the representativeness of the CPCSSN population, and to identify people with diabetes in the CPCSSN data and describe their emergency department visits and hospital admissions.	Patient sex, patient age, patient postal code, health conditions table, risk factor table, medications, lab table, referral table, disease case
Williamson et al.[54]	2017	Manual review of electronic medical records as a reference standard for case definition development: a validation study	Determine whether the processed and standardized data contained with the CPCSSN database might function as a reference standard for case definition validation.	Health conditions table, risk factor table, billing, encounter diagnosis, medications, lab table, allergy table, exam (height, weight, bp) , medical procedures, referral table
Bharathi et al.[55]	2015	Validation of identification of Bell's Palsy cases in Canadian primary Care EMR data - a pilot study	Assess the feasibility of finding Bell's palsy patients in large EMR databases and measure the level of completeness and accuracy of recorded data in Canadian Family Practice EMRs about Bell's palsy	Patient sex, patient age, health conditions table, billing, encounter diagnosis, medications, exam (height, weight, bp) , referral table
Aliarzadeh et al.[56]	2016	Hypertension screening and follow-up in children and adolescents in a Canadian primary care population sample: a retrospective cohort study	Identify current rates of pediatric hypertension screening and follow-up in Canada. Examine patient and provider characteristics associated with rates of blood pressure screening.	Patient sex, patient age, provider demographics, exam (height, weight, bp) , site table
Biro et al.[57]	2016	Prevalence of toddler, child and adolescent overweight and obesity derived from primary care electronic medical records: an observational study	Determine the prevalence of childhood overweight and obesity using objective measures derived from primary care electronic medical records.	Patient sex, patient age, billing, exam (height, weight, bp)
Biro et al.[58]	2016	Utility of linking primary care electronic medical records with Canadian census	Test the feasibility of extracting full postal code from primary care EMRs and	Patient sex, patient age, patient postal

Author	Year	Study Title	Study Objectives	Tables/variables used
		data to study the determinants of chronic disease: an example based on socioeconomic status and obesity	linking this with area-level measures of the environment to demonstrate how such a linkage could be used to examine the determinants of disease.	code, exam (height, weight, bp)
Cave et al.[59]	2016	Development of a validated algorithm for the diagnosis of paediatric asthma in electronic medical records	Develop and validate a case definition of asthma in children 1–17 who consult family physicians.	Patient sex, patient age, health conditions table, billing, encounter diagnosis, medications, lab table, exam (height, weight, bp) , referral table
Drummond et al.[60]	2016	Prevalence and management of dementia in primary care practices with electronic medical records: a report from the Canadian Primary Care Sentinel Surveillance Network	Describe the prevalence and management of dementia in a community-dwelling sample using electronic medical record (EMR) data.	Patient sex, patient age, patient postal code, health conditions table, billing, encounter diagnosis, medications, exam (height, weight, bp) , disease case
Greiver et al.[61]	2016	Implementation of data management and effect on chronic disease coding in a primary care organisation: A parallel cohort observational study	Implement data management activities in a community-based primary care organisation and to evaluate the effects on coding for chronic conditions.	Patient sex, patient age, provider demographics, health conditions table, billing, encounter diagnosis, EMR
Lukewich et al.[62]	2016	Association between registered nurse staffing and management outcomes of patients with type 2 diabetes within primary care: a cross-sectional linkage study	Examine the relation between primary care delivery models that incorporate registered nurses and clinical outcomes of patients with type 2 diabetes.	Patient sex, patient age, health conditions table, billing, encounter diagnosis, lab table, exam (height, weight, bp) , network id
Perveen et al.[63]	2016	Performance analysis of data mining classification techniques to predict diabetes	Classify patients with diabetes mellitus using diabetes risk factors following the adaboost and bagging ensemble techniques using J48 (c4.5) decision tree as a base learner along with standalone data mining technique J48	Patient sex, patient age, health conditions table, billing, encounter diagnosis, lab table, exam (height, weight, bp)
Queenan et al.[64]	2016	Herpes zoster infection in people with diabetes in Canadian primary care practice.	Determine the burden of herpes zoster infection in people with diabetes in CPCSSN.	Patient sex, patient age, health conditions table, billing, encounter diagnosis, disease case
Queenan et al.[65]	2016	Representativeness of patients and providers in the Canadian Primary Care Sentinel Surveillance Network: a cross-sectional study.	Determine how representative the data for patients and primary care practitioners in the CPCSSN are when compared with the Canadian population.	Patient sex, patient age, patient postal code, provider demographics, site table

Author	Year	Study Title	Study Objectives	Tables/variables used
Rosella et al.[66]	2016	Prevention and Screening for Type 2 Diabetes	Determine what interventions are likely to be effective in reducing the incidence of Type 2 Diabetes and its medical complications in the adult population of Newfoundland and Labrador.	Patient sex, patient age, health conditions table, billing, encounter diagnosis, lab table, network id, disease case
Singer et al.[67]	2016	Data quality of electronic medical records in Manitoba: do problem lists accurately reflect chronic disease billing diagnoses?	Determine problem list completeness related to chronic diseases in electronic medical records (EMRs) and explore clinic and physician factors influencing completeness.	Provider demographics, health conditions table, billing, site table
Singian et al.[68]	2016	Using Canadian Primary Care Sentinel Surveillance Network data to examine depression in patients with a diagnosis of Parkinson disease: a retrospective cohort study	Describe demographic and health characteristics of patients with Parkinson disease and examine sex differences in antidepressant prescriptions for those with comorbid depression using electronic medical records.	Patient sex, patient age, patient postal code, health conditions table, risk factor table, billing, encounter diagnosis, medications, lab table, exam (height, weight, bp)
Aref-Eshghi et al.[69]	2015	Low density lipoprotein cholesterol control status among Canadians at risk for cardiovascular disease: findings from the Canadian Primary Care Sentinel Surveillance Network Database.	Determine the prevalence of uncontrolled LDL-C in patients with high cardiovascular disease (CVD) risks across Canada and to examine its related factors.	Patient sex, patient age, patient postal code, health conditions table, risk factor table, billing, encounter diagnosis, medications, lab table, exam (height, weight, bp) , disease case
Asghari et al.[70]	2015	Single and mixed dyslipidaemia in Canadian primary care settings: findings from the Canadian primary care sentinel surveillance network database.	Describe the prevalence of various single and mixed dyslipidaemia within the Canadian population in a primary care setting.	Patient sex, patient age, patient postal code, health conditions table, risk factor table, billing, encounter diagnosis, medications, lab table, exam (height, weight, bp)
Asghari et al.[71]	2015	Does the prevalence of dyslipidemias differ between Newfoundland and the rest of Canada? Findings from the electronic medical records of the Canadian Primary Care Sentinel Surveillance Network	Compare the prevalence of dyslipidemia in the NL population with the rest of Canada.	Patient sex, patient age, patient postal code, health conditions table, risk factor table, billing, encounter diagnosis, medications, lab table, exam (height, weight, bp) , disease case
Barber et al.[72]	2015	Data discipline in electronic medical records: Improving smoking status documentation with a standardized intake tool and process.	Evaluate the transformation in smoking status documentation after implementing a standardized intake tool as part of a primary care smoking cessation program.	Risk factor table

Author	Year	Study Title	Study Objectives	Tables/variables used
Birtwhistle et al.[73]	2015	Prevalence and management of osteoarthritis in primary care: an epidemiologic cohort study from the Canadian Primary Care Sentinel Surveillance Network	Describe the epidemiology of osteoarthritis in patients aged 30 years and older using electronic medical records (EMRs) in a Canadian primary care population.	Patient sex, patient age, patient postal code, health conditions table, risk factor table, medications, exam (height, weight, bp) , disease case
Coleman et al.[74]	2015	From patient care to research: a validation study examining the factors contributing to data quality in a primary care electronic medical record database	Examine and describe the validity of chronic disease case definition algorithms and factors affecting data quality in a primary care EMR database.	Patient sex, patient age, health conditions table, risk factor table, billing, encounter diagnosis, medications, lab table, allergy table, exam (height, weight, bp) , network id, medical procedures, site table, referral table, disease case
Farahani et al.[75]	2015	Exploring the distribution of prescription for sulfonylureas in patients with type 2 diabetes according to cardiovascular risk factors within a Canadian primary care setting	Explore the distribution of established atherosclerotic CV disease and CV risk factors amongst patients with diabetes on an SU using a Canadian primary care dataset for the 2013 calendar year.	Patient sex, patient age, health conditions table, risk factor table, billing, encounter diagnosis, medications, lab table, disease case
Garies et al.[76]	2015	Using EMR data to evaluate a physician-developed lifestyle plan for obese patients in primary care	Use primary care electronic medical records (EMRs) to evaluate the effects of a lifestyle intervention delivered to obese patients compared with obese patients who did not receive the intervention.	Patient sex, patient age, patient postal code, health conditions table, billing, encounter diagnosis, medications, lab table, exam (height, weight, bp)
Godwin et al.[77]	2015	Prevalence and management of hypertension in primary care practices with electronic medical records: a report from the Canadian Primary Care Sentinel Surveillance Network	Report on the prevalence and management of hypertension based on data from the Canadian Primary Care Sentinel Surveillance Network	Patient sex, patient age, patient postal code, health conditions table, risk factor table, billing, medications, exam (height, weight, bp) , disease case
Green et al.[78]	2015	Chronic obstructive pulmonary disease in primary care: an epidemiologic cohort study from the Canadian Primary Care Sentinel Surveillance Network	Determine the prevalence of physician-diagnosed COPD in primary care practices, and the degree of comorbidity with other chronic conditions, and to assess patterns of medication prescribing.	Patient sex, patient age, patient postal code, risk factor table, medications, exam (height, weight, bp) , disease case
Greiver et al.[79]	2015	Are we asking patients if they smoke? missing information on tobacco use in Canadian Electronic Medical Records.	Examine the recording of tobacco use in Canadian electronic medical records.	Patient sex, patient age, provider demographics, risk factor table, network id, site table, disease case

Author	Year	Study Title	Study Objectives	Tables/variables used
Mashayekhi et al.[80]	2015	Evaluating the performance of the Framingham Diabetes Risk Scoring Model in Canadian electronic medical records	Evaluate the performance of the Framingham Diabetes Risk Scoring Model in a Canadian population.	Patient sex, patient age, lab table, exam (height, weight, bp) , disease case
Morkem et al.[81]	2015	A Canadian Primary Care Sentinel Surveillance Network study evaluating antidepressant prescribing in Canada from 2006 to 2012.	Evaluate the prescribing patterns of antidepressants (ADs) by primary care providers to youth, adults, and seniors.	Patient sex, patient age, medications
Nicholson et al.[82]	2015	Examining the symptom of fatigue in primary care: a comparative study using electronic medical records	Examine the prevalence and impact of patients presenting with fatigue in primary care, using the only known electronic database in Canada to capture patient-reported symptoms.	Patient sex, patient age, billing, encounter diagnosis, network id, referral table, disease case
Nicholson et al.[83]	2015	Examining the prevalence and patterns of multimorbidity in Canadian primary healthcare: a methodologic protocol using a national electronic medical record database	Measure the point prevalence and clusters of multimorbidity among adult primary health care (PHC) patients; and examine the natural history and changing burden of multimorbidity over time among adult PHC patients.	Patient sex, patient age, patient postal code, billing, encounter diagnosis, site table
Ogunleye et al.[84]	2015	Depression, diabetes and multimorbidity: results from the Northern Alberta Primary Care Research Network Data	Evaluate the association between diabetes and depression.	Patient sex, patient age, health conditions table, billing, encounter diagnosis, medications, lab table, exam (height, weight, bp) , network id, disease case
Rigobon et al. [85]	2015	Adult obesity prevalence in primary care users: An exploration using Canadian Primary Care Sentinel Surveillance Network (CPCSSN) data.	Examine the feasibility of using electronic medical records within the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) for obesity surveillance in Canada	Patient sex, patient age, exam (height, weight, bp)
Aliarzadeh et al.[86]	2014	Association between socio-economic status and hemoglobin A1c levels in a Canadian primary care adult population without diabetes	Study the association between neighborhood-level SES and Hgb A1c in a primary care population without diabetes.	Patient sex, patient age, patient postal code, health conditions table, lab table, exam (height, weight, bp) , disease case
Greiver et al.[87]	2014	Prevalence and epidemiology of diabetes in Canadian primary care practices: a report from the Canadian Primary Care Sentinel Surveillance Network	Describe the epidemiology of diabetes in this Canadian sample.	Patient sex, patient age, patient postal code, health conditions table, billing, medications, lab table, disease case

Author	Year	Study Title	Study Objectives	Tables/variables used
Maddocks et al.[88]	2014	Identifying new referrals from FPs using EMRs	Examine patterns of referral to specialist physicians.	Patient sex, patient age, health conditions table, referral table
Williamson et al.[89]	2014	Validating the 8 CPCSSN case definitions for chronic disease surveillance in a primary care database of electronic health records	Develop and validate case definitions and case-finding algorithms used to identify 8 common chronic conditions in primary care: chronic obstructive pulmonary disease (COPD), dementia, depression, diabetes, hypertension, osteoarthritis, parkinsonism, and epilepsy.	Health conditions table, billing, encounter diagnosis, medications, lab table, exam (height, weight, bp) , network id, disease case
Williamson et al.[90]	2014	CPCSSN's Role in improving pharmacovigilance	Examine completeness of CPCSSN adverse event data.	Patient sex, patient age, billing, encounter diagnosis, medications
Wong et al.[91]	2014	The diagnosis of depression and its treatment in Canadian primary care practices: an epidemiological study	Describe the prevalence of the diagnosis among men and women, patient characteristics and drug treatment in patients diagnosed with depression in the primary care setting in Canada.	Patient sex, patient age, patient postal code, health conditions table, risk factor table, billing, encounter diagnosis, medications, exam (height, weight, bp) , disease case
Oake et al.[92]	2013	Prevalence of dyslipidemia in Newfoundland adults: approaches to estimation using electronic medical records	Assess the prevalence of dyslipidemia in NL using Canadian Primary Care Sentinel Surveillance Network (CPCSSN) EMR data. Develop an algorithm that will provide a more accurate estimation of dyslipidemia using EMR data.	Patient sex, patient age, patient postal code, health conditions table, billing, encounter diagnosis, medications, lab table, disease case
Greiver et al.[93]	2013	The Canadian Primary Care Sentinel Surveillance Network (CPCSSN): Developing a method to estimate practice denominators for a national Canadian electronic medical record database	Develop and test a method to calculate primary care practice denominators (the 'population at risk' being studied).	Patient sex, patient age, provider demographics
Kadhim-Saleh et al.[94]	2013	Validation of the diagnostic algorithms for 5 chronic conditions in the Canadian Primary Care Sentinel Surveillance Network (CPCSSN): A Kingston Practice-based Research Network (PBRN)	Assess the validity of electronic medical records–based diagnostic algorithms for 5 chronic conditions.	Patient sex, patient age, health conditions table, billing, encounter diagnosis, medications, lab table, exam (height, weight, bp) , medical procedures, referral table, disease case
Torti et al.[95]	2013	Documenting alcohol use in primary care in Alberta	Determine the proportion of patients with alcohol use documented in EMRs and the number of ways physicians	Risk factor table

Author	Year	Study Title	Study Objectives	Tables/variables used
			documented alcohol use in order to describe the patterns of documentation.	
Greiver et al.[96]	2012	Who are your patients with diabetes?: EMR case definitions in the Canadian Primary Care setting	Demonstrate how CPCSSN provides an opportunity to develop and validate case definitions across Canada using many different EMRs.	Health conditions table, billing, medications, lab table, disease case
Greiver et al.[97]	2011	Diabetes screening with hemoglobin A1c prior to a change in guideline recommendations: prevalence and patient characteristics	Determine the provision of tests in non-diabetic patients age 19 or over, patients age 45 and over (eligible for routine diabetes screening), the annual change in the rate of this screening test, and the patient characteristics associated with the provision of Hgb A1c screening.	Patient sex, patient age, patient postal code, health conditions table, risk factor table, billing, encounter diagnosis, medications, lab table, exam (height, weight, bp) , network id

References

1. Singer A, Kosowan L, Loewen S, Spitoff S, Greiver M, Lynch J. Who is asked about alcohol consumption? A retrospective cohort study using a national repository of Electronic Medical Records. *Prev Med Rep*. 2021;22:101346. doi:10.1016/j.pmedr.2021.101346

2. Barber D, Morkem R, Dalgarno N, et al. Patients eligible and referred for bariatric surgery in southeastern Ontario: Retrospective cohort study. *Can Fam Physician*. 2021;67(1):e31-e40. doi:10.46747/cfp.6701e31

3. Singer AG, Kosowan L, Soller L, et al. Prevalence of Physician-Reported Food Allergy in Canadian Children. *J Allergy Clin Immunol Pract*. 2021;9(1):193-199. doi:10.1016/j.jaip.2020.07.039

4. Black JE, Terry AL, Lizotte DJ. Development and evaluation of an osteoarthritis risk model for integration into primary care health information technology. *Int J Med Inform*. 2020;141:104160. doi:10.1016/j.ijmedinf.2020.104160

5. Ross S, Fast H, Garies S, et al. Pelvic floor disorders in women who consult primary care clinics: development and validation of case definitions using primary care electronic medical records. *CMAJ Open*. 2020;8(2):E414-E419. doi:10.9778/cmajo.20190145

6. Puzhko S, Schuster T, Barnett TA, et al. Evaluating Prevalence and Patterns of Prescribing Medications for Depression for Patients With Obesity Using Large Primary Care Data (Canadian Primary Care Sentinel Surveillance Network). *Front Nutr*. 2020;7:24. doi:10.3389/fnut.2020.00024

7. Miyagishima R, Drummond N, Carroll L, Hopper T, Garies S, Williamson T. Validation of a case definition for speech and language disorders: In community-dwelling older adults in Alberta. *Can Fam Physician*. 2020;66(3):e107-e114.

8. Marrie RA, Kosowan L, Singer A. Management of diabetes and hypertension in people with multiple sclerosis. *Mult Scler Relat Disord*. 2020;40:101987. doi:10.1016/j.msard.2020.101987

9. Cave AJ, Soos B, Gillies C, Drummond N, Pham ANQ, Williamson T. Validating a case definition for adult asthma in primary care electronic medical records. *NPJ Prim Care Respir Med*. 2020;30(1):24. doi:10.1038/s41533-020-0181-3

10. Bang F, Ehsani B, McFaull S, et al. Surveillance of concussion-related injuries using electronic medical records from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN): a proof-of-concept. *Can J Public Health*. 2020;111(2):193-201. doi:10.17269/s41997-019-00267-4

11. Bello AK, Ronksley PE, Tangri N, et al. Prevalence and Demographics of CKD in Canadian Primary Care Practices: A Cross-sectional Study. *Kidney Int Rep*. 2019;4(4):561-570. doi:10.1016/j.ekir.2019.01.005

12. Garies S, Hao S, McBrien K, 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. doi:10.1001/jamanetworkopen.2019.0406

13. Rigobon AV, Kalia S, Nichols J, et al. Impact of the Diabetes Canada Guideline Dissemination Strategy on the Prescription of Vascular Protective Medications: A Retrospective Cohort Study, 2010-2015. *Diabetes Care*. 2019;42(1):148-156. doi:10.2337/dc18-0935

14. Kosowan L, Wicklow B, Queenan J, Yeung R, Amed S, Singer A. Enhancing Health Surveillance: Validation of a Novel Electronic Medical Records-Based Definition of Cases of Pediatric Type 1 and Type 2 Diabetes Mellitus. *Can J Diabetes*. 2019;43(6):392-398. doi:10.1016/j.jcjd.2019.02.005

15. O'Neill B, Kalia S, Aliarzadeh B, et al. Agreement between primary care and hospital diagnosis of schizophrenia and bipolar disorder: A cross-sectional, observational study using record linkage. *PLoS One*. 2019;14(1):e0210214. doi:10.1371/journal.pone.0210214

16. Perveen S, Shahbaz M, Keshavjee K, Guergachi A. Metabolic Syndrome and Development of Diabetes Mellitus: Predictive Modeling Based on Machine Learning Techniques. *IEEE Access*. 2019;7:1365-1375. doi:10.1109/ACCESS.2018.2884249

17. Phillips SP, Jiang M, Lakkadghatwala R, Wang S. Assessing wellness in the well-child check: What about social and emotional development? *Can Fam Physician*. 2019;65(3):e113-e120.

18. Greiver M, Kalia S, Voruganti T, 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):e024970. doi:10.1136/bmjopen-2018-024970

19. Abu-Ashour W, Twells LK, Valcour JE, Gamble JM. Diabetes and the occurrence of infection in primary care: a matched cohort study. *BMC Infect Dis*. 2018;18(1):67. doi:10.1186/s12879-018-2975-2

20. Drummond N, Taylor M, Garies S, et al. Developing and implementing linked electronic medical record and administrative data in primary care practice for diabetes in Alberta. *International Journal of Population Data Science*. 2018;3(4). doi:10.23889/ijpds.v3i4.969

21. Drummond N, McCleary L, Freiheit E, et al. Antidepressant and antipsychotic prescribing in primary care for people with dementia. *Can Fam Physician*. 2018;64(11):e488-e497.

22. Ehsani-Moghaddam B, Queenan JA, MacKenzie J, Birtwhistle RV. Mucopolysaccharidosis type II detection by Naïve Bayes Classifier: An example of patient classification for a rare disease using electronic medical records from the Canadian Primary Care Sentinel Surveillance Network. *PLoS One*. 2018;13(12):e0209018. doi:10.1371/journal.pone.0209018

23. Bartlett G. Antidepressant Prescription Practices among Primary Health Care Providers for Patients with Diabetes Mellitus. *CRDOJ*. 2017;2(4). doi:10.19080/CRDOJ.2017.02.555593

24. Kalia S, Greiver M, Zhao X, et al. Would you like to add a weight after this blood pressure, doctor? Discovery of potentially actionable associations between the provision of multiple screens in primary care. *J Eval Clin Pract*. 2018;24(2):423-430. doi:10.1111/jep.12877

25. Hurd J, Pike A, Knight J, et al. Health and health service use of very elderly Newfoundlanders. *Can Fam Physician*. 2018;64(10):e453-e461.

26. Williamson T, Aponte-Hao S, Mele B, et al. Developing and Validating a Primary Care EMR-based Frailty Definition using Machine Learning. *Int J Popul Data Sci*. 2020;5(1):1344. doi:10.23889/ijpds.v5i1.1344

27. Gagnon J, Lussier MT, MacGibbon B, Daskalopoulou SS, Bartlett G. The Impact of Antidepressant Therapy on Glycemic Control in Canadian Primary Care Patients With Diabetes Mellitus. *Front Nutr*. 2018;5:47. doi:10.3389/fnut.2018.00047

28. Legacy Drug-Prescribing Patterns in Primary Care - PubMed. Accessed April 29, 2022. https://pubmed.ncbi.nlm.nih.gov/30420366/

29. McAlister FA, Garrison S, Kosowan L, Ezekowitz JA, Singer A. Use of Direct Oral Anticoagulants in Canadian Primary Care Practice 2010-2015: A Cohort Study From the Canadian Primary Care Sentinel Surveillance Network. *J Am Heart Assoc*. 2018;7(3). doi:10.1161/JAHA.117.007603

30. Weaver C, Garies S, Williamson T, McBrien K, Peng M. Association rule mining to identify potential under-coding of conditions in the problem list in primary care electronic medical records. *International Journal of Population Data Science*. 2018;3(4). doi:10.23889/ijpds.v3i4.622

31. Perveen S, Shahbaz M, Keshavjee K, Guergachi A. A Systematic Machine Learning Based Approach for the Diagnosis of Non-Alcoholic Fatty Liver Disease Risk and Progression. *Sci Rep*. 2018;8(1):2112. doi:10.1038/s41598-018-20166-x

32. Katz A, Wong S, Williamson T, Taylor C, Peterson S. Identification of Frailty using EMR and Admin data: A complex issue. *International Journal of Population Data Science*. 2018;3(4). doi:10.23889/ijpds.v3i4.832

33. Queenan JA, Farahani P, Ehsani-Moghadam B, Birtwhistle RV. 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):465-469. doi:10.1016/j.jcjd.2017.10.060

34. Reyes RR, Parker G, Garies S, et al. Team-based comanagement of diabetes in rural primary care. *Can Fam Physician*. 2018;64(8):e346-e353.

35. Greiver M, Sullivan F, Kalia S, et al. Agreement between hospital and primary care on diagnostic labeling for COPD and heart failure in Toronto, Canada: a cross-sectional observational study. *npj Prim Care Resp Med*. 2018;28(1):1-8. doi:10.1038/s41533-018-0076-8

36. Singer A, Fanella S, Kosowan L, et al. Informing antimicrobial stewardship: factors associated with inappropriate antimicrobial prescribing in primary care. *Fam Pract*. 2018;35(4):455-460. doi:10.1093/fampra/cmx118

37. Singer A, Kosowan L, Katz A, Jolin-Dahel K, Appel K, Lix LM. Prescribing and testing by primary care providers to assess adherence to the Choosing Wisely Canada recommendations: a retrospective cohort study. *CMAJ Open*. 2018;6(4):E603-E610. doi:10.9778/cmajo.20180053

38. Bello AK, Ronksley PE, Tangri N, et al. A national surveillance project on chronic kidney disease management in Canadian primary care: a study protocol. *BMJ Open*. 2017;7(8):e016267. doi:10.1136/bmjopen-2017-016267

39. Wong S, Katz A, Williamson T, Peterson S, Taylor C, McGrail K. Can Linked Electronic Medical Record and Administrative Data Help Us Identify Those Living With Frailty? *International Journal of Population Data Science*. 2018;3(4). doi:10.23889/ijpds.v3i4.829

40. Abrams EM, Singer AG, Lix L, Katz A, Yogendran M, Simons FER. Adherence with epinephrine autoinjector prescriptions in primary care. *Allergy Asthma Clin Immunol*. 2017;13:46. doi:10.1186/s13223-017-0218-5

41. O'Brien MA, Sullivan F, Carson A, Siddiqui R, Syed S, Paszat L. Piloting electronic screening forms in primary care: findings from a mixed methods study to identify patients eligible for low dose CT lung cancer screening. *BMC Fam Pract*. 2017;18(1):95. doi:10.1186/s12875-017-0666-5

42. 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-e503.

43. Coons MJ, Greiver M, Aliarzadeh B, et al. Is glycemia control in Canadians with diabetes individualized? A cross-sectional observational study. *BMJ Open Diabetes Res Care*. 2017;5(1):e000316. doi:10.1136/bmjdrc-2016-000316

44. Loo CKJ, Greiver M, Aliarzadeh B, Lewis D. Association between neighbourhood walkability and metabolic risk factors influenced by physical activity: a cross-sectional study of adults in Toronto, Canada. *BMJ Open*. 2017;7(4):e013889. doi:10.1136/bmjopen-2016-013889

45. Lix L, Singer A, Katz A, Yogendran M, Al-Azazi S. Chronic Disease Case Definitions for Electronic Medical Records: A Canadian Validation Study: IJPDS (2017) Issue 1, Vol 1:212 Proceedings of the IPDLN Conference (August 2016). *International Journal of Population Data Science*. 2017;1(1). doi:10.23889/ijpds.v1i1.232

46. 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. doi:10.5210/ojphi.v9i1.7648

47. Morkem R, Patten S, Queenan J, Barber D. Recent Trends in the Prescribing of ADHD Medications in Canadian Primary Care. *J Atten Disord*. 2020;24(2):301-308. doi:10.1177/1087054717720719

48. Morkem R, Williamson T, Patten S, et al. Trends in antidepressant prescribing to children and adolescents in Canadian primary care: A time-series analysis. *Pharmacoepidemiol Drug Saf*. 2017;26(9):1093-1099. doi:10.1002/pds.4240

49. Oake J, Aref-Eshghi E, Godwin M, et al. Using Electronic Medical Record to Identify Patients With Dyslipidemia in Primary Care Settings: International Classification of Disease Code Matters From One Region to a National Database. *Biomed Inform Insights*. 2017;9:1178222616685880. doi:10.1177/1178222616685880

50. Aref-Eshghi E, Oake J, Godwin M, et al. Identification of Dyslipidemic Patients Attending Primary Care Clinics Using Electronic Medical Record (EMR) Data from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) Database. *J Med Syst*. 2017;41(3):45. doi:10.1007/s10916-017-0694-7

51. Singer A, Kroeker AL, Yakubovich S, Duarte R, Dufault B, Katz A. Data quality in electronic medical records in Manitoba: Do problem lists reflect chronic disease as defined by prescriptions? *Can Fam Physician*. 2017;63(5):382-389.

52. Ryan BL, Shadd J, Maddocks H, Stewart M, Thind A, Terry AL. Methods to Describe Referral Patterns in a Canadian Primary Care Electronic Medical Record Database: Modelling Multilevel Count Data. *J Innov Health Inform*. 2017;24(4):888. doi:10.14236/jhi.v24i4.888

53. Birtwhistle R, Green ME, Frymire E, 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-E564. doi:10.9778/cmajo.20170017

54. Williamson T, Miyagishima RC, Derochie JD, Drummond N. Manual review of electronic medical records as a reference standard for case definition development: a validation study. *CMAJ Open*. 2017;5(4):E830-E833. doi:10.9778/cmajo.20170077

55. Bharathi R, Sullivan F, Aliarzadeh B, Greiver M. Validation of Identification of Bell's palsy Cases in Canadian Primary Care EMR Data-A Pilot Study. *Annals of Otolaryngology and Rhinology*. 2015;3:1082.

56. Aliarzadeh B, Meaney C, Moineddin R, et al. Hypertension screening and follow-up in children and adolescents in a Canadian primary care population sample: a retrospective cohort studystudy. *CMAJ Open*. 2016;4(2):E230-E235. doi:10.9778/cmajo.20150016

57. Biro S, Barber D, Williamson T, Morkem R, Khan S, Janssen I. Prevalence of toddler, child and adolescent overweight and obesity derived from primary care electronic medical records: an observational study. *CMAJ Open*. 2016;4(3):E538-E544. doi:10.9778/cmajo.20150108

58. Biro S, Williamson T, Leggett JA, et al. Utility of linking primary care electronic medical records with Canadian census data to study the determinants of chronic disease: an example based on socioeconomic status and obesity. *BMC Med Inform Decis Mak*. 2016;16:32. doi:10.1186/s12911-016-0272-9

59. Cave AJ, Davey C, Ahmadi E, 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. doi:10.1038/npjpcrm.2016.85

60. Drummond N, Birtwhistle R, Williamson T, Khan S, Garies S, Molnar F. 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):E177-184. doi:10.9778/cmajo.20150050

61. Greiver M, Wintemute K, Aliarzadeh B, et al. Implementation of data management and effect on chronic disease coding in a primary care organisation: A parallel cohort observational study. *J Innov Health Inform*. 2016;23(3):843. doi:10.14236/jhi.v23i3.843

62. Lukewich J, Edge DS, VanDenKerkhof E, Williamson T, Tranmer J. Association between registered nurse staffing and management outcomes of patients with type 2 diabetes within primary care: a cross-sectional linkage study. *CMAJ Open*. 2016;4(2):E264-270. doi:10.9778/cmajo.20150113

63. Perveen S, Shahbaz M, Guergachi A, Keshavjee K. Performance Analysis of Data Mining Classification Techniques to Predict Diabetes. *Procedia Computer Science*. 2016;82:115-121. doi:10.1016/j.procs.2016.04.016

64. Queenan J, Farahani P, Khan S, Birtwhistle R. Herpes Zoster Infection in People with Diabetes in Canadian Primary Care Practice. *Canadian Journal of Diabetes*. 2016;40:S17. doi:10.1016/j.jcjd.2016.08.051

65. Queenan JA, Williamson T, Khan S, 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. doi:10.9778/cmajo.20140128

66. Rosella, L., Bornstein, S., Mackey, S., Grignon, M. (2016). Prevention and Screening for Type 2 Diabetes in Newfoundland and Labrador. St. John's, NL: Newfoundland & Labrador Centre for Applied Health Research, Memorial University

67. 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):1107-1112. doi:10.1093/jamia/ocw013

68. Singian KRP, Price M, Bungay V, Wong ST. Using Canadian Primary Care Sentinel Surveillance Network data to examine depression in patients with a diagnosis of Parkinson disease: a retrospective cohort study. *CMAJ Open*. 2016;4(3):E417-E423. doi:10.9778/cmajo.20160052

69. Aref-Eshghi E, Leung J, Godwin M, et al. Low density lipoprotein cholesterol control status among Canadians at risk for cardiovascular disease: findings from the Canadian Primary Care Sentinel Surveillance Network Database. *Lipids Health Dis*. 2015;14:60. doi:10.1186/s12944-015-0056-8

70. Asghari S, Aref-Eshghi E, Godwin M, Duke P, Williamson T, Mahdavian M. Single and mixed dyslipidaemia in Canadian primary care settings: findings from the Canadian primary care sentinel surveillance network database. *BMJ Open*. 2015;5(12):e007954. doi:10.1136/bmjopen-2015-007954

71. Asghari S, Aref-Eshghi E, Hurley O, et al. Does the Prevalence of Dyslipidemias Differ between Newfoundland and the Rest of Canada? Findings from the Electronic Medical Records of the Canadian Primary Care Sentinel Surveillance Network. *Front Cardiovasc Med*. 2015;2:1. doi:10.3389/fcvm.2015.00001

72. Barber D, Williamson T, Biro S, et al. Data discipline in electronic medical records: Improving smoking status documentation with a standardized intake tool and process. *Can Fam Physician*. 2015;61(12):e570-576.

73. Birtwhistle R, Morkem 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):E270-275. doi:10.9778/cmajo.20150018

74. Coleman N, Halas G, Peeler W, Casaclang 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. doi:10.1186/s12875-015-0223-z

75. Farahani P, Khan S, Oatway M, Dziarmaga A. Exploring the Distribution of Prescription for Sulfonylureas in Patients with Type 2 Diabetes According to Cardiovascular Risk Factors Within a Canadian Primary Care Setting. *J Popul Ther Clin Pharmacol*. 2015;22(3):e228-236.

76. 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):e225-231.

77. Godwin M, Williamson T, Khan S, 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. doi:10.9778/cmajo.20140038

78. Green ME, Natajaran N, O'Donnell DE, 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. doi:10.9778/cmajo.20140040

79. Greiver M, Aliarzadeh B, Meaney C, 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):264-268. doi:10.1016/j.amepre.2015.01.005

80. Mashayekhi M, Prescod F, Shah B, Dong L, Keshavjee K, Guergachi A. Evaluating the performance of the Framingham Diabetes Risk Scoring Model in Canadian electronic medical records. *Can J Diabetes*. 2015;39(2):152-156. doi:10.1016/j.jcjd.2014.10.006

81. Morkem R, Barber D, Williamson T, Patten SB. A Canadian Primary Care Sentinel Surveillance Network Study Evaluating Antidepressant Prescribing in Canada From 2006 to 2012. *Can J Psychiatry*. 2015;60(12):564-570. doi:10.1177/070674371506001207

82. Nicholson K, Stewart M, Thind A. Examining the symptom of fatigue in primary care: a comparative study using electronic medical records. *J Innov Health Inform*. 2015;22(1):235-243. doi:10.14236/jhi.v22i1.91

83. Nicholson K, Terry AL, Fortin M, Williamson T, Bauer M, Thind A. Examining the prevalence and patterns of multimorbidity in Canadian primary healthcare: a methodologic protocol using a national electronic medical record database. *J Comorb*. 2015;5:150-161. doi:10.15256/joc.2015.5.61

84. Ogunleye A, Manca D, Sharma A, Campbell-Scherer D. Depression, Diabetes and Multi-Morbidity: Results from the Northern Alberta Primary Care Research Network Data. *Canadian Journal of Diabetes*. 2015;39:S27-S28. doi:10.1016/j.jcjd.2015.01.113

85. Rigobon AV, Birtwhistle R, Khan S, 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):e283-289. doi:10.17269/cjph.106.4508

86. Aliarzadeh B, Greiver M, Moineddin R, et al. Association between socio-economic status and hemoglobin A1c levels in a Canadian primary care adult population without diabetes. *BMC Fam Pract*. 2014;15:7. doi:10.1186/1471-2296-15-7

87. Greiver M, Williamson T, Barber D, 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):179-185. doi:10.1016/j.jcjd.2014.02.030

88. Maddocks H, Ryan BL, Shadd J, Terry A, Chevendra V. Identifying new referrals from FPs using EMRs. *Can Fam Physician*. 2014;60(10):949.

89. Williamson T, Green ME, Birtwhistle R, 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):367-372. doi:10.1370/afm.1644

90. Williamson T, Lévesque L, Morkem R, Birtwhistle R. CPCSSN's role in improving pharmacovigilance. *Can Fam Physician*. 2014;60(7):678.

91. Wong ST, Manca D, Barber D, et al. The diagnosis of depression and its treatment in Canadian primary care practices: an epidemiological study. *CMAJ Open*. 2014;2(4):E337-E342. doi:10.9778/cmajo.20140052

92. Oake J, Asghari S, Godwin M, Collins K, Aubrey K. Prevalence of Dyslipidemia in Newfoundland Adults: Approaches to Estimation Using Electronic Medical Records. *J Epidemiol Community Health*. 2013;67(4):e1-e1. doi:10.1136/jech-2013-202386.4

93. Greiver M, Williamson T, Bennett TL, et al. Developing a method to estimate practice denominators for a national Canadian electronic medical record database. *Fam Pract*. 2013;30(3):347-354. doi:10.1093/fampra/cms083

94. Kadhim-Saleh A, Green M, Williamson T, Hunter D, Birtwhistle R. Validation of the diagnostic algorithms for 5 chronic conditions in the Canadian Primary Care Sentinel Surveillance Network (CPCSSN): a Kingston Practice-based Research Network (PBRN) report. *J Am Board Fam Med*. 2013;26(2):159-167. doi:10.3122/jabfm.2013.02.120183

95. 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.

96. Greiver M, Keshavjee K, Martin K, Aliarzadeh B. Who are your patients with diabetes? *Can Fam Physician*. 2012;58(7):804.

97. Greiver M, Aliarzadeh B, Moineddin R, Meaney C, Ivers N. Diabetes screening with hemoglobin A1c prior to a change in guideline recommendations: prevalence and patient characteristics. *BMC Fam Pract*. 2011;12:91. doi:10.1186/1471-2296-12-91

Appendix B: Algorithms to Identify Patients with the Test Conditions.

Condition	Definition – Gold Standard [Reference Standard]	ICD9 Diagnosis Code in Billing Table
Diabetes Mellitus	Diagnosis on the problem list or insulin prescription, or >=2 oral anti-diabetic agents (excluding Metformin), or 1+ Metformin prescriptions, or >=2 abnormal plasma glucose tests >=3 months apart (fasting plasma glucose>=7.0 mmol/L; casual(random) plasma glucose >=11.1 mmol/L; 2 hour plasma glucose in a 75-g oral glucose tolerance test >=11.1 mmol/L), or HbA1c>=6.5%[1] Dataset C: Note – results were not tested for diabetic patients because the dataset does not distinguish between random and casual blood glucose laboratory tests, and the threshold values for the two tests are different.	250
Hypertension	Patients >=18 years with: (Diagnosis on problem list OR 3+ systolic BP readings above the healthy adult threshold OR 3+ Diastolic BP readings above the health adult threshold - >=140 systolic, or >=90 diastolic) OR (1+ Prescriptions for a diuretic OR 2+ anti-hypertensive oral medications)	401
Hypothyroidism[2]	Diagnosis on problem list or Prescription for: Hypothyroidism medications or TSH level>10	244
Asthma	Patients <18 years with (Diagnosis on problem list AND 1+ medications) OR (2+ types of medications – a controller and relief medication)	493
Obesity	Patients >=18 years with: (Diagnosis on the Problem list as obesity OR obese) OR (BMI on Problem List >=30) OR (Calculated BMI 30+ (use height and weight))	278
Urinary Tract Infection	Diagnosis on the Problem list AND 1+ oral antibiotic medications	595

[1] Adapted Gold Standard Definition for Diabetes Case Ascertainment as defined by Harris SB, Glazier RH, Tompkins JW, Wilton AS, Chevendra V, Stewart MA et al.: Investigating concordance in diabetes diagnosis between primary care charts (electronic medical records) and health administrative data: a retrospective cohort study. BMC Health Serv Res 2010, 10: 347.

[2]Definition adapted from Hassey (2001). Hassey A, Gerrett D, Wilson A. A survey of validity and utility of electronic patient records in a general practice. BMJ 2001;322:1401-1405.

Appendix B was originally published in:

Terry, A.L., Stewart, M., Cejic, S. *et al.* A basic model for assessing primary health care electronic medical record data quality. *BMC Med Inform Decis Mak* **19,** 30 (2019). https://doi.org/10.1186/s12911-019-0740-0

Appendix C: Medications Used in Condition Identification

Condition	Medications
Diabetes Mellitus	One of: Insulin, Humulin, Iletin, Pork, detemir, glargine, Levemir, Lantus, Novolin, protamine, Humalog, aspart, novomix, novorapid, glulisine OR At least two of: Acarbose, Glucobay, saxagliptin, Onglyza, sitagliptin, Janumet, Januvia, nateglnide, Repaglinide, gluconorm, chlorpropamide, Sulfonylureas, Gliclazide, diamicron, glimepiride, avandaryl, glyburide, Diabeta, Tolbutamide, pioglitazone, Actos, Rosiglitazone, Avandamet, Avandaryl, Avandia OR One or more of: METFORMIN, AVANDAMET, GLUCHOPHAGE, GLUMETZA, GLYCON, JANUMET, Glimepiride, rosiglitazone, Avandaryl
Hypertension	One or more of: labetalol, doxazosin, cardura, prazosin, apo-prazo, nu-prazo, terazosin, hytrin, phentolamine, pindolol, viskazide, visken, nadolol, propranolol, acebutolol, monitan, rhotral, sectral, atenolol, tenoretic, tenormin, bisoprolol, monocor, esmolol, brevibloc, metoprolol, clonidine, catapres, dixarit, methyldopa, aldomet, supres, amlodipine, caduet, calcium channel, norvasc, twynsta, felodipine, plendil, renedil, nifedipine, adalat, nifed , diltiazem, verapamil, aliskiren, benazepril, ace inhibitor, lotensin, captopril, apo-capto, capoten, captril, cilazapril, inhibace, enalapril, enalaprilat, vasotec, fosinopril, monopril, lisiniopril, prinivil, prinzide, zestoretic, zestril, perindopril, coversyl, quinapril, accupril, accuretic, ramipril, altace, trandolapril, mavik, candesartan, atacand, eprosartan, teveten, irebesartan, avalide, avapro, losartan, cozaar, hyzaar, olmesartan, olmetec, telmisartan, micardis, twynsta, valsartan, diovan, trandolapril, verapamil, tarka, amlodipine/atorvastatin, Caduet, methyldopa/HCT, methyldopa HCT, methyldopa/hydro, supres, aliskiren fumarate/hydro, aliskiren fumarate/HCT, aliskiren fumarate HCT, rasilez HCT, rasilez/HCT, diazoxide, hyperstat I.V., proglycem, epoprostenol, flolan, hydralazine hydrochloride, hydralazine/hydro, hydralazine/HCT, Apo®-Hydralazine, Apresoline, Novo-Hylazin, Nu-Hydral, minoxidil, Apo®-Gain, Loniten, Rogaine, sodium nitroprusside, Nipride OR 2 or more of: amiloride, amilzide, spironolactone, aldactazide, triamterene, triazide, dyazide, chlorthalidone, atenidone, tenoretic, HYDROCHLOROTHIAZIDE, ACCURETIC, ALDACTAZIDE, ALTACE, AMILZIDE, CILAZAPRIL, APO-HYDRO, TRIAZIDE, ATACAND, AVALIDE, IRBESARTAN, DIOVAN, DYAZIDE, HYDRODIURIL, HYZAAR, INDERIDE, INHIBACE, MICARDIS, LISINOPRIL, NU-HYDRO, TRIAZIDE, OLMETEC, PRINZIDE, RASILEZ, VALSARTAN, SER-AP-ES, ENALAPRIL, TRIAMTERENE, TEVETEN, THIAZIDE, TIMOLIDE, VASERETIC, VISKAZIDE, ZESTORETIC, COVERSYL PLUS, LOZIDE, INDAPAMIDE, cilazapril/hydro, inhibace plus, cilazapril/HCT, enalapril/HCT, enalapril/HYDRO, enalapril HCT, vaseretic, lisinopril/HCT, lisinopril HCT, lisinopril/hydro, prinzide, zestoretic, perindopril/HCT, perindopril HCT, perindopril/hydro, quinapril/HCT, quinapril HCT, quinapril/hydro, accuretic, ramipril/HCT, ramipril HCT, ramipril/hydro, altace HCT, altace/HCT, candesartan/HCT, candesartan HCT, candesartan/HYDRO, atacand plus, eprosartan/HCT, eprosartan HCT, eprosartan/HYDRO, teveten plus, irbesartan/HCT, irbesartan HCT, irbesartan/HYDRO, avalide, losartan/HCT, losartan HCT, losartan/HYDRO, hyzaar, olmesartan/HCT, olmesartan HCT, olmesartan/hydro, olmetec plus, telmisartan/HCT, telmisartan HCT, telmisartan/hydro, micardis plus, valsartan/HCT, valsartan HCT, valsartan/hydro,

Condition	Medications
	diovan/HCT, diovan HCT, diovan-HCT, ATENOLOL/CHLOR, ATENOLOL CHLOR, APO-ATENIDONE, APO ATENIDONE, TENORETIC
Hypothyroidism	One or more of: levothyroxine, Eltroxin, Levotec, Synthroid, Thyroid, liothyronine, Cytomel,
Asthma	One or more of: epinephrine HCL, Adrenalin, Vaponefrin, formoterol fumarate, Foradil, oxeze, symbicort, zenhale, salbutamol, salvent, terbutaline, bricanyl, isoproteronol, orciprenaline, alupent, ipratropium, ipravent, atrovent, tiotropium bromide, spiriva, BECLOMETHASONE, BECLODISK, BECLOFORTE, BECLOVENT, BECONASE, CORTICOSTEROIDS: INHALED, MYLAN-BECLO, PROPADERM, QVAR, RIVANASE, VACENASE, budesonide, pulmicort, rhinocort, symbicort, ciclesonide, alvesco, fluticasone, cutivate, flonase, flovent, CROMOGLYCATE, CROMOLYN, INTAL SPINCAPS, NALCROM, opticrom, RHINARIS, FENOTEROL HYDROBROMIDE/IPRATROPIUM, duovent, formoterol fumarate dihydrate/budenoside, symbicort, IPRATROPIUM BROMIDE/SALBUTAMOL, COMBIVENT, MYLAN-COMBO, RATIO-IPRA, salmeterol xinafoate/flut, advair, cortisone acetate, serevent, barriere-hc, cipro hc, hydrocortisone sodium succinate, solu-cortef, methylprednisolone sodium succinate, Solu-Medrol, prednisolone sodium phosphate, inflamase, pediapred, vasocidin, prednisone, triamcinolone, aristocort, montelukast, Singulair, zafirlukast, Accolate, aminophy, Phylloc, Theophy, oxtriphylline, Choledyl, Rouphylline, THEOPHYLLINE, APO-THEO, QUIBRON, RATIO-THEO-BRONC, SLO-BID, THEO-DUR, Theolair, UNIPHYL, omalizumab OR (one or more of: formoterol fumarate, Foradil, oxeze, symbicort, zenhale, salbutamol, salvent, terbutaline, bricanyl, isoproteronol, orciprenaline, alupent, ipratropium, ipravent, atrovent, tiotropium bromide, spiriva, BECLOMETHASONE, BECLODISK, BECLOFORTE, BECLOVENT, BECONASE, CORTICOSTEROIDS: INHALED, MYLAN-BECLO, PROPADERM, QVAR, RIVANASE, VACENASE, budesonide, pulmicort, rhinocort, symbicort, ciclesonide, alvesco, fluticasone, cutivate, flonase, flovent, CROMOGLYCATE, CROMOLYN, INTAL SPINCAPS, NALCROM, opticrom, RHINARIS, FENOTEROL HYDROBROMIDE/IPRATROPIUM, duovent, formoterol fumarate dihydrate/budenoside, symbicort, IPRATROPIUM BROMIDE/SALBUTAMOL, COMBIVENT, MYLAN-COMBO, RATIO-IPRA, salmeterol xinafoate/flut, advair, cortisone acetate, serevent, barriere-hc, cipro hc, hydrocortisone sodium succinate, solu-cortef, methylprednisolone sodium succinate, Solu-Medrol, prednisolone sodium phosphate, inflamase, pediapred, vasocidin, prednisone, triamcinolone, aristocort, montelukast, Singulair, zafirlukast, Accolate, aminophy, Phylloc, Theophy, oxtriphylline, Choledyl, Rouphylline, THEOPHYLLINE, APO-THEO, QUIBRON, RATIO-THEO-BRONC, SLO-BID, THEO-DUR, Theolair, UNIPHYL, omalizumab AND One or more of: epinephrine HCL, Adrenalin, Vaponefrin)
Urinary Tract Infection	One or more of: cephalex, Keflex, cefaclor, ceclor, cefuroxime, ceftin, cefixime, Suprax, Proloprim, Trimethoprim , methenamine, mandelamine, nitrofurantoin, amoxi, Hp-PAC, Aclavulanate, ampicillin, ampicin, nu-ampi, cipro oral, cipro XL, apo-ciproflox, cipro tab, Fluoroquinolones , ciprofloxacin, levofloxacin, Levaquin, norflox, noroxin tab, CO-TRIMOXAZOLE, TRIMETHOPRIM/SULF, TRIMETHOPRIM SULF, SULFATRIM, BACTRIM ROCHE, NU-COTRIMOX, SEPTRA DS, SEPTRA TAB, SULFAMETHOXAZOLE-TRIMETHOPRIM, SULFAMETHOXAZOLE TRIMETHOPRIM

Appendix D: Calculation of Sensitivity and Positive Predictive Value

Reference Standard vs. Index Test (Billing Code) 2X2 Table

	Reference standard criteria present	Reference standard criteria absent
Billing code present	True Positive	False Positive
Billing code absent	False Negative	True Negative

Sensitivity = True positive/ (True positive + False negative)

Sensitivity = # of patients who have both the reference standard criteria and the billing code/ (# of patients who have the reference standard criteria)

Positive Predictive Value = True positive / (True positive + False Positive)

Positive Predictive Value = # of patients who have both the reference standard criteria and the billing code/ (# of patients who have the billing code)

Example: Calculated sensitivity and positive predictive value for urinary tract infection (UTI) where the reference standard criteria was having “URINARY TRACT INFECTION” or “UTI” in the ‘DiagnosisText_calc’ column of the ‘HealthCondition’ table OR “595” or “595.x” (x could be any number) in the ‘DiagnosisCode_calc’ column of the ‘HealthCondition’ table

AND 1 or more of: [medications omitted. See Appendix C] In the ‘Name_calc’ column of the ‘Medication’ table

	Reference standard criteria present	Reference standard criteria absent	Total
Billing code present	318	-	18,748
Billing code absent	-	-	-
Total	1,787	-	-

Sensitivity= # of patients who have both the reference standard criteria for UTI and the billing code for UTI/ (# of patients who have the reference standard criteria for UTI)

Sensitivity =318/1787

Sensitivity =17.8%

Positive predictive value= # of patients who have both the reference standard criteria for UTI and the billing code for UTI/ (# of patients who have the billing code for UTI)

Positive Predictive Value =318/18748

Positive Predictive Value =1.7%

Appendix E: Ethics Approval

Date: 14 June 2020

To: Dr. Amanda Terry

Project ID: 115903

Study Title: Development of a Multi-Factorial Data Quality Score for Primary Care Electronic Medical Records

Application Type: HSREB Initial Application

Review Type: Delegated

Meeting Date / Full Board Reporting Date: 16/Jun/2020

Date Approval Issued: 14/Jun/2020

REB Approval Expiry Date: 14/Jun/2021

Dear Dr. Amanda Terry

The Western University Health Science Research Ethics Board (HSREB) has reviewed and approved the above mentioned study as described in the WREM application form, as of the HSREB Initial Approval Date noted above. This research study is to be conducted by the investigator noted above. All other required institutional approvals must also be obtained prior to the conduct of the study.

Documents Approved:

Document Name	Document Type	Document Date
CPCSSN-ERD-v4.0.4	Other Data Collection Instruments	14/Jun/2020
Primary Care EMR Data Quality Score - Protocol	Protocol	14/Jun/2020

Documents Acknowledged:

Document Name	Document Type	Document Date
References for Rationale	References	14/Jun/2020

No deviations from, or changes to, the protocol or WREM application should be initiated without prior written approval of an appropriate amendment from Western HSREB , except when necessary to eliminate immediate hazard(s) to study participants or when the change(s) involves only administrative or logistical aspects of the trial.

REB members involved in the research project do not participate in the review, discussion or decision.

The Western University HSREB operates in compliance with, and is constituted in accordance with, the requirements of the TriCouncil Policy Statement: Ethical Conduct for Research Involving Humans (TCPS 2); the International Conference on Harmonisation Good Clinical Practice Consolidated Guideline (ICH GCP); Part C, Division 5 of the Food and Drug Regulations; Part 4 of the Natural Health Products Regulations; Part 3 of the Medical Devices Regulations and the provisions of the Ontario Personal Health Information Protection Act (PHIPA 2004) and its applicable regulations. The HSREB is registered with the U.S. Department of Health & Human Services under the IRB registration number IRB 00000940.

Please do not hesitate to contact us if you have any questions.

Sincerely,

Daniel Wyzynski, Research Ethics Coordinator, on behalf of Dr. Philip Jones, HSREB Vice-Chair

Note: This correspondence includes an electronic signature (validation and approval via an online system that is compliant with all regulations).

Western Research

Date: 21 June 2021

To: Dr. Amanda Terry

Project ID: 115903

Study Title: Development of a Multi-Factorial Data Quality Score for Primary Care Electronic Medical Records

Application Type: Continuing Ethics Review (CER) Form

Review Type: Delegated

REB Meeting Date: 06/July/2021

Date Approval Issued: 21/Jun/2021 13:30

REB Approval Expiry Date: 14/Jun/2022

***Ethics Approval Lapse:** June 15 - June 21, 2021*

Dear Dr. Amanda Terry,

The Western University Research Ethics Board has reviewed the application. This study, including all currently approved documents, has been re-approved until the expiry date noted above.

REB members involved in the research project do not participate in the review, discussion or decision.

Western University REB operates in compliance with, and is constituted in accordance with, the requirements of the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans (TCPS 2); the International Conference on Harmonisation Good Clinical Practice Consolidated Guideline (ICH GCP); Part C, Division 5 of the Food and Drug Regulations; Part 4 of the Natural Health Products Regulations; Part 3 of the Medical Devices Regulations and the provisions of the Ontario Personal Health Information Protection Act (PHIPA 2004) and its applicable regulations. The REB is registered with the U.S. Department of Health & Human Services under the IRB registration number IRB 00000940.

Please do not hesitate to contact us if you have any questions.

Sincerely,

The Office of Human Research Ethics

Note: This correspondence includes an electronic signature (validation and approval via an online system that is compliant with all regulations).