CPCSSN Data Quality

An Opportunity for Enhancing Canadian Primary Care Data

April 2023

Rachael Morkem[1,2], Ayat Salman[1,2],

Chad Herman[1,2], Richa Shah[3], Sabrina Wong[3,4], David Barber[1,2]

[1] Canadian Primary Care Sentinel Surveillance Network, Queen's University
[2] Department of Family Medicine, Queen's University
[3] Centre for Health Services Research and School of Nursing, University of British Columbia
[4]National Institute of Nursing Research, National Institutes of Health

Acknowledgements

The authors wish to recognize contributions to CPCSSN from the following:

Practice Based Research and Learning Network directors and scientists

Kris Aubrey-Bassley (Memorial University), David Barber (Queen's University), Simone Dahrouge (University of Ottawa), Kimberly Fairman (Institute of Circumpolar Research), Stephanie Garies (University of Calgary), Mathew Grandy (Dalhousie University), Michelle Griever (University of Toronto), Marie-Thérèse Lussier (Université de Montréal), Donna Manca (University of Alberta), Dee Mangin (McMaster University), Kerry McBrien (University of Calgary), Aude Motulsky (Université de Montréal), Andrew Pinto (University of Toronto), Jennifer Rayner (Western University), Claude Richard (Université de Montréal), Alex Singer (University of Manitoba), Amanda Terry (Western University), Tyler Williamson (University of Calgary), Brianne Wood (Northern School of Medicine), Sabrina Wong (National Institute for Nursing Research) and Barbara Zelek (Northern School of Medicine).

Data managers

Kris Adamczyk, Babak Aliarazadeh, Charles Bruntz, Tao Chen, Michael Cummings, Brian Forst, Andy Gibb, Lorne Kinsella, Jennifer Lawson, Bill Peeler, Sara Sabri, Lena Schofield, Bedredine Sayah, Boglarka Soos, Matt Taylor, Rick Truant and Don White.

Research assistants/associates

Alexis Aomreore, Aashka Bhatt, Allison Boileau, Lanting Cheng, Kimberly Duerkson, Keri Harvey, Leanne Kosowan, Jennifer Lawson, Aimie Lee, Jodi Lees, Himisara Marasinghe, Walid Nacher, Fazle Sharior, Rabiya Siddiqui, and Rebecca Theal.

Reviewers

Simone Dahrouge, Michael Cummings, Annalise Schamuhn, Cliff Lindeman and Stephanie Garies.

Publication information

How to cite this publication

Morkem R, Salman A, Herman C, Shah R, Barber D. CPCSSN Data and Information Quality: An Opportunity for Enhancing Canadian Primary Care Data. Kingston, ON. Canadian Primary Care Sentinel Surveillance Network; April 2023.

About CPCSSN

The Canadian Primary Care Sentinel Surveillance Network (CPCSSN) is an independent not-for-profit university-based consortium with an international reputation as a trusted source of primary care electronic medical record (EMR) data. Established in 2008, CPCSSN has developed a pan-Canadian primary care EMR data repository. CPCSSN has successfully built trusting relationships between primary care clinicians and researchers over the past 15 years. As of 2022, CPCSSN consisted of a network of 13 community-based primary care research and learning networks based in eight Canadian provinces (British Columbia, Alberta, Manitoba, Ontario, Quebec, Nova Scotia, Newfoundland) and one territory (Northwest Territories). CPCSSN is also working to increase its representativeness across Canada with the development of a new network in Saskatchewan. CPCSSN draws on technological expertise to securely extract EMR data from primary care practices and includes close to 1,500 participating primary care providers and approximately 2 million patients. CPCSSN applies standardized ontologies and terminologies to transform data from various EMR vendors into a common data schema. CPCSSN goes to great lengths to protect privacy and was recognized in 2013 with a privacy innovation award for being a leader in maintaining the security of health information.

CPCSSN is supported through a diverse array of funding including peer-reviewed grants from federal agencies such as the Canadian Institutes of Health Research and the Public Health Agency of Canada, as well as other profit and not-for-profit organizations. CPCSSN is also supported with in-kind and direct funding support from the Canadian universities that host the regional networks.

Contents

Executive Summary

Purpose

The purpose of this report is to assess the quality of the CPCSSN database and to provide key learnings and recommendations to inform users so that they may assess its fitness for use.

We use a data quality framework from the National Quality Assurance Framework (NQAF) to define five dimensions of quality. Each dimension is assessed using specific and measurable evidence-based indicators.

Findings

- Overall, the CPCSSN database has reasonable data quality for epidemiological and populated-based research.
- The CPCSSN database captures a wide spectrum of data types (e.g., diagnostic codes, medications, labs, exams), which provides access to current and past patient health records and information on healthcare delivery.
- The quality of data is high in terms of element agreement, validity, distributions of clinical parameters, and comparison to other data sources. The element presence (completeness) indicator highlights the extensive work CPCSSN has done to create coded, standardized information.
- The comparability and coherence of the database is perhaps where CPCSSN has the poorest data quality. The indicator for this dimension revealed a great degree of variation in the use of common ICD9 codes, medications, and labs at each site, within each province, and by EMR type.
- While CPCSSN has developed a large library of supplementary and explanatory information to educate and inform users about the database this information needs to be more accessible and available to users of the data.

Key Recommendations

- To remain relevant, CPCSSN must continue to work closely with users, clients, and stakeholders.
- Fitness-for-use would be increased by the development of standardized methodology that would enable users to link various clinical data elements (diagnoses, prescriptions, labs) to an encounter (visit).
- CPCSSN operations should continue to develop cleaning and processing tools to reduce the missingness in coded fields as much as possible. Higher priority items include expanding the list of labs that are extracted and coded; and improving the coding of medication metrics (e.g., duration, strength).

- It is recommended that users request identification of site, EMR and province so that clustering at these levels can be accounted for in the analysis. Furthermore, in some contexts, researchers may want to consider different analytical approaches depending on the EMR or province source.
- The accessibility and clarity of the CPCSSN data could be improved by making the supporting information accessible, available, and more clearly understood. This could include improvements to the CPCSSN website design to ensure it has intuitive navigation, concise content, strategic use of visuals and usable forms.
- We recommend the creation of a training module or user-friendly data dictionary and resource guide, which could include a shared repository of code for data preparation, for researchers and analysts to guide them through CPCSSN and its data holdings, from acquisition to analysis.
- Lastly, this report highlights the need for a sophisticated dynamic tool that would allow a researcher to easily evaluate whether the CPCSSN dataset is suitable for their specific purpose. Such a data utility evaluation could not only be useful for users to identify whether the database meets the minimum requirements for their purpose but would also enable CPCSSN operations to identify where to invest resources in data quality improvements.

Introduction

A learning healthcare system (LHS) is a value-based healthcare system that strives to achieve the best possible outcome at the lowest cost. There are four elements that characterize an LHS (a) core values, (b) pillars and accelerators, (c) processes, and (d) outcomes. A fundamental requirement for establishing an LHS is the generation and management of robust practice-based data.[1,2] Such data drives research, surveillance, and quality improvement and can lead to substantial changes in practice.[1] For this to happen, the clinical data in a database must be of good quality, or at the very least, of known quality.[2,3] Easy access to reliable population-level clinical data enables evidence-driven health system transformation and is a vital component to operationalize an LHS.[1]

To ensure data quality, it is important to conduct regular data quality assessments. While data cleaning is often completed prior to analyses, data quality assessments are rarely performed.[4] A data quality assessment of clinical data derived from electronic medical records (EMRs) is especially crucial, as these data are complex and often contain many unstructured elements. In addition, there are inherent challenges in the extraction and consolidation of clinical data from EMR software due to inconsistent or non-existent EMR specifications and standards.[5] Interpreting and making sense of the data can also be challenging, due to differing patterns of care and widely varying documentation habits of healthcare professionals across organizations and jurisdictions (i.e., provinces). This heterogeneity coupled with the non-random human errors that may occur across multiple dimensions of the data need to be reported so that the data can be effectively evaluated, and to ensure the knowledge gleaned from this clinical information is reliable and accurate.[3]

The Canadian Primary Care Sentinel Surveillance Network (CPCSSN) is Canada's first EMR surveillance system and aims to improve primary healthcare delivery outcomes across the country, while also facilitating innovation and excellence in primary healthcare research. It is essential that information within this database be of good quality as it is an important resource to improve health care delivery in Canada.

Data quality must be consistently defined in context of its production and use.[6] To effectively assess quality, an overarching framework is needed to provide a clear picture of the concepts that define quality and that gives context for quality concerns, activities, and initiatives.[3,7] The National Quality Assurance Framework (NQAF) has developed guidelines to help organizations formulate and operationalize national quality frameworks.[3,7] The NQAF template defines quality in terms of the following components: (1) **relevance**, (2) **accuracy and reliability**, (3) **timeliness and punctuality**, (4) **coherence and comparability**, and (5) **accessibility and clarity**. These measures of data quality have been adopted by other Canadian organizations that hold clinical and statistical data (e.g., Statistics Canada, Canadian Institute for Health Information).[3,7-9] These five data quality dimensions form the foundation of the data quality assessment of the CPCSSN database detailed in this report.

The goal of this report is to present an assessment of CPCSSN data quality and to provide key learnings and recommendations to inform users so that they may assess its fitness for use. This is an important endeavour, as primary health care is the foundation of the healthcare system and the data held within EMRs can provide invaluable insight into patient health and health care delivery.[10]

CPCSSN Data Quality Framework

The following dimensions were adapted from the NQAF framework and the Canadian Institute for Health Information (CIHI) Information Quality Framework and will be used to describe and assess CPCSSN data quality.[7-9]

RELEVANCE

The degree to which the information meets users' current and potential future needs.

ACCURACY AND RELIABILITY

The degree to which the information correctly and consistently describes the phenomena it was designed to measure.

COMPARABILITY AND COHERENCE

The degree to which information is comparable over time and across jurisdictions, produced using common standards and methods, and can be combined with other sources.

TIMELINESS AND PUNCTUALITY

Timeliness refers to how quickly information is made available after the end of the reference period; punctuality refers to whether information is delivered on the announced dates.

ACCESSIBILITY AND CLARITY

The degree to which information, including supplementary and explanatory information and metadata, is easily obtainable and clearly presented, in a way that can be understood.

Indicators

Choosing specific and measurable indicators for each of the five dimensions is vital to effectively evaluate the quality of an EMR-derived dataset.[3] In addition to the guidelines and framework provided by the NQAF, extensive reviews provide recommendations on systematic, statistically based methods of data quality assessments.[4,7,8,11-14] The indicators chosen to evaluate the data quality of the CPCSSN database were selected based on up-to-date literature and was adapted to the specific characteristics of the CPCSSN data.[3,11-14]

Coded CPCSSN Data

A general understanding of how CPCSSN data are processed is important in constructing useful indicators of data quality. The basic data processing pipeline that CPCSSN follows is:

1. the data are extracted from individual clinics, direct identifiers are removed, and each patient is assigned a unique random CPCSSN patient ID;
2. each clinic's EMR data is mapped and transformed into the CPCSSN data structure;
3. the data is then cleaned and converted into standard ontologies, e.g., Anatomical Therapeutical Chemical (ATC) classification, Logical Observation Identifiers Names and Codes (LOINC), and International Classification of Diseases (ICD-9)
4. validated disease case detection algorithms are applied to the database to classify patients as having target conditions; and
5. the data from each network are merged to create a single research-grade database in the CPCSSN pan-Canadian repository.

Following extraction from the EMR, the data are cleaned and coded using a sophisticated array of processing tools, some of which are based on natural language processing (NLP) and machine learning (ML) methods. In addition to cleaning and standardizing the raw (original) data extracted from the EMR, CPCSSN processing also creates derived data elements. This includes the application of validated and published case detection algorithms that use a range of data elements to identify a patient as having a health condition (e.g., Diabetes Mellitus).[15-22]

CPCSSN's processing (cleaning, coding, and transformation) is a continuous and adaptive process. It responds to broadening information requirements and changes to health systems and technologies over each data extraction cycle. CPCSSN holds data from ten different EMR products across seven different provinces. Subsequently, the CPCSSN database is heterogeneous, and this complexity was an important consideration in how this data quality assessment was undertaken. This report focusses mainly on evaluating the data quality of the coded data within the CPCSSN database; however, to provide a more complete picture, we provide metrics on some of the original (or raw) data that has not yet been cleaned or transformed into a useable format.

Methods

This report used data from the 2022-Q2 data extraction, which includes records up until June 30, 2022. For some indicators we evaluated all records within the database, while for other indicators we defined a denominator - that is, a specified patient cohort. Where a defined denominator was necessary, this report used a two-year contact group (2YCG), as this represents patients who are more actively engaged in the health system and are more likely to have up to date documentation.[23,24] As there were likely disruptions in care that occurred due to the Covid-19 pandemic we chose a 2018-2019 2YCG to estimate an active patient population. Furthermore, for some indicators we evaluated all records associated with the 2018-2019 2YCG (up until June 30, 2022), but where the pandemic may have impacted a given

data quality indicator (for example, measurement and documentation of a BMI), we only included records as of December 31, 2019.

This report does not include an assessment of the representativeness of the CPCSSN database to the Canadian population, as this information has been previously published.[25]

Glossary

Dates: When a provider records information within the EMR the date that the record is entered is automatically documented within the EMR system. When raw data is extracted from the EMR and mapped to the CPCSSN schema, these 'system recorded' date values populate a field called 'DateCreated' within each CPCSSN table. For some types of data, such as medications, labs or health conditions, there is also an associated date field (e.g., the Medication table has a start and stop date) where the provider has recorded in the EMR the actual date of service or condition onset. These additional dates are specific to each table.

Coded versus Uncoded: If a variable is coded it means there are algorithms that have taken the raw, unstructured data and transformed it into a standardized format. Uncoded data has uncertain usability.

Null: The variable is missing any text or numeric strings. Note that some fields may become populated with a value after processing all data within the EMR (i.e., if a diagnosis code is present the coding and cleaning processes will populate the diagnosis name using the code information).

2010 Start Date: Only records ≥ 2010 were included in this data quality evaluation. As of 2010, EMRs had been widely adopted and in use by primary care clinics, and software and documentation patterns had improved to a consistent level.

Geographic Measures: Due to ethics and privacy policies, not all CPCSSN-contributing networks submit full postal codes. As such, these data (from which other geographic measures, such as social and material deprivation and rural or urban status, are derived) are not always available within the CPCSSN database.

Data Quality Assessment of the CPCSSN database

In the sections that follow, the CPCSSN database is evaluated on the five core data quality dimensions using defined indicators. Each quality dimension, its indicators, and the findings from the application of that indicator, are described.

1. RELEVANCE

This dimension describes the degree to which the information meets existing and target users' current and potential needs. This refers to whether data elements that are required by users are produced and are useful, particularly the extent to which the data concepts, definitions and classifications correspond to user needs.[7,9] This data quality dimension is measured using two indicators: (a) Network Participation; and (b) Use and Access.

(A) NETWORK PARTICIPATION

Description: Indicates the size and jurisdictions of the networks that participate in CPCSSN.

Calculation description: the number of networks that participate in CPCSSN, the size (number of providers and patients) of each network, and location (province) of each network.

Type of measure: descriptive.

Findings

The data that comprises the CPCSSN database comes from 13 PBLRNs that collect EMR data from 1,444 primary care providers operating out of 268 unique clinics ('sites'). The database contains primary care data on 1,819,192 patients and the median number of patients per provider (practice) is 752. Table 1 describes the data contributions from each PBRLN.

Table 1. CPCSSN PBLRNs

	Region	Patients	Providers	EMR Count	Site Count
CPCSSN	Pan-Canadian	1,819,192	1,444	9	268
Practice Based Research and Learning Networks (PBLRNs)					
1	British Columbia	134,557	101	4	22
2	Southern Alberta	247,040	181	4	27
3	Northern Alberta	106,455	86	4	16
4	Manitoba	267,036	249	2	47
5	Southwestern Ontario	*No data available at this time*			
6	Western Ontario	89,803	56	2	4
7	Central Ontario	584,360	389	3	98
8	Southeastern Ontario	217,705	148	3	14
9	Eastern Ontario	2,097	3	1	2
10	Northern Ontario	27,509	44	3	9
11	Quebec	46,493	130	2	11
12	Nova Scotia	85,001	64	1	17
13	Newfoundland	11,136	21	1	1

(B) Use and Access

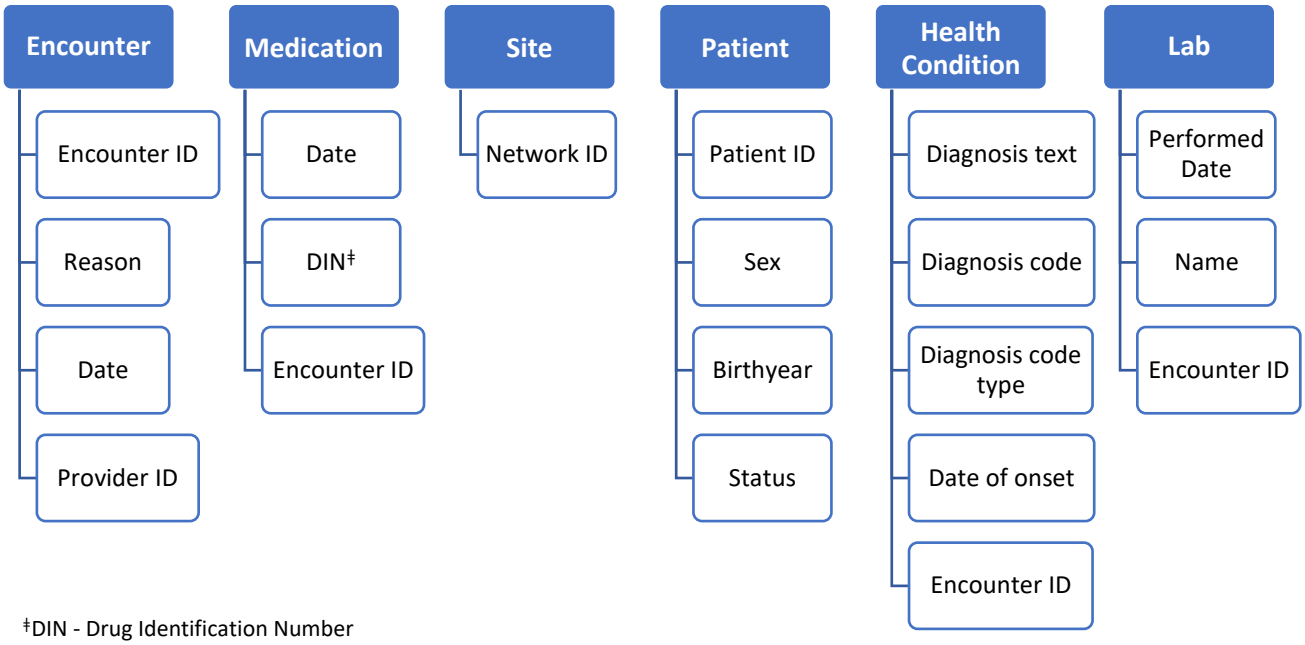
Description: A summary of past and current users of the database and the most often requested data elements.

Calculation description: A list of the top 20 data elements requested by researchers; a description of the number, funding, and type of projects that used the CPCSSN database in CPCSSN's first ten years (2008-2018) and last four years (2019-2022).

Type of measure: percentages, descriptive.

Findings

Figure 1. Top 20 Data Elements Requested



‡DIN - Drug Identification Number

The top twenty data elements requested by researchers in the last five years (2017-2022) are listed in Figure 1. Unsurprisingly, this list reveals that data users are interested in a variety of patient demographic and clinical elements including birthyear, sex, diagnoses, medications, and labs. Another common variable that is requested is the Encounter Identification (ID), which is an identification number that links various data elements (medications, labs, diagnoses) to a specific encounter (visit). Ideally this data element would be a key part of the architecture of an EMR and could be easily extracted and incorporated into the CPCSSN database. Unfortunately, there is inconsistent use of minimum or standard data elements and this lack of EMR content standards means that the Encounter ID is an unreliable data element. Currently, researchers are advised to use associated dates to link different data elements around a patient encounter or visit. This is an important data element and CPCSSN could increase its relevancy by developing a calculated (derived) encounter identification data element that could be incorporated into the CPCSSN database.

An evaluation of the types of projects that CPCSSN data is used for reveals that there is wide variation in data requirements that exist amongst users of CPCSSN data. The database is primarily used for research and disease surveillance but is also a resource for quality improvement activities related to patient care. Since its inception in 2008, the pan-Canadian database has been accessed and used for 110 projects, of which most focused on chronic disease, data science and epidemiologic methods, while a smaller number of projects were data quality or patient quality improvement initiatives. In the first five years of the database's existence (2008-2013), 50% to 60% of the data access and use requests came from academic clinicians and researchers associated with one of the CPCSSN PBRLNs (considered 'internal'). In the years since, CPCSSN has become increasingly known as a source of data that does not exist elsewhere and this is reflected in the increase in partnerships between academic researchers and clinicians with industry and government in the last five years.

Many of CPCSSN's operational activities are focused on developing and improving CPCSSN's data and on producing relevant, high-quality information products. Understanding who is using the data and what their data needs are is a central component to producing a relevant data product. As such, CPCSSN continues to engage with stakeholders and data users to inform CPCSSN's priorities on the data and on improving its quality.

2. ACCURACY AND RELIABILITY

Accuracy reflects the degree to which the information correctly and consistently describes the phenomena it was designed to measure (e.g., the degree of closeness of estimates to true values). This can be characterized by estimating sampling and non-sampling error, which is decomposed into bias (systematic error) and variance (random error).[7,9] Reliability is concerned with whether the data consistently (over time or across variables) measures the reality that they are designed to represent.[7,9] There are five indicators that will be used to measure this data quality dimension: (a) Element Presence, (b) Element Agreement, (c) Data Source Agreement, (d) Distribution Comparison, and (e) Validity Check.[4,13,14]

(A) Element Presence

Description: There are common data elements that are expected to be present, or 'not null' (i.e., a data entry exists).[4] This indicator presents the percent present for the main variables within the tables that comprise the CPCSSN database.

Calculation description: The common data elements (variables) within each table of the database will be assessed to determine if the information in that variable has been standardized (coded vs uncoded), and to describe the completeness of the uncoded and coded versions (% present). Data elements with a < 10% complete are red, 10-69% complete are yellow, and ≥ 70% complete are green. Lastly, the presence of key covariates (that are often requested for epidemiological studies) will be evaluated: social and material deprivation, rural or urban location, body mass index (BMI) and smoking status.

Type of measure: Percentages, descriptive.

Findings

In the CPCSSN data, there are common data elements expected to be 'not null' based on the presence of a record. Table 2 lists common data elements within each table of the CPCSSN database, whether these data elements have undergone standardization into a calculated (coded) field, and the percent present, or completeness, for the uncoded (original/raw data from EMRs) and coded (standardized) version of that data element. It is important to note that some coded data elements are calculated using data from two or more original fields, so that it may be possible for the coded field to have a higher completeness than its uncoded counterpart. Data elements that contain dates are not presented as coded or uncoded, as this information is extracted from the EMR in a standard format.

Table 2. Element Presence

CPCSSN DATA TABLE	DATA ELEMENT (VARIABLE)	CODED (Y/N)	UNCODED % PRESENT	CODED % PRESENT
ALLERGY INTOLERANCE	Name	Y	91.11%	50.96%
	Code	Y	3.90%	50.96%
	Category	N	50.28%	0.00%
	Severity	Y	69.74%	32.90%
	Status	Y	69.68%	53.82%
	Reaction Type	N	46.67%	0.00%
	Start Date	–	-	59.09%
	Stop Date	–	-	0.13%
	Date Created	–	-	92.80%
BILLING	Service Code	Y	99.81%	95.25%
	Diagnosis Text	Y	46.39%	74.90%
	ICD9 Code	Y	79.49%	74.90%
	Service Date	–	-	99.98%
	Date Created	–	-	99.49%
ENCOUNTER	Encounter Type	Y	41.01%	45.45%
	Reason	N	54.49%	0.00%
	Encounter Date	–	-	100.00%
	Date Created	–	-	96.33%
ENCOUNTER DIAGNOSIS	Diagnosis Code	Y/N	79.23%	75.64%
	Diagnosis Text	Y/N	94.06%	75.64%
	Date Created	–	-	100.00%
EXAM	Exam Name	Y	94.88%	5.12%
	Exam Result	Y	94.88%	5.12%
	Date Created	–	100.00%	
FAMILY HISTORY	Diagnosis Text	Y	100.00%	34.68%
	Diagnosis Code	Y	16.38%	34.68%

	Relationship	Y		65.56%
	Relationship side	Y	54.78%	45.82%
	Relationship Degree	Y		64.73%
	Age at Onset	Y	7.34%	
	Vital Status	Y	4.70%	5.71%
	Was Cause of Death	Y		59.76%
	Date Created	–	99.15%	
HEALTH CONDITION	Diagnosis Code	Y/N	42.79%	45.32%
	Diagnosis Text	Y/N	99.67%	45.32%
	Status	Y	90.92%	100.00%
	Date of Onset	–	28.27%	
	Date Created	–	99.42%	
MEDICATION	ATC Code	Y	44.63%	96.75%
	Name	Y	99.93%	96.75%
	Reason	N	7.89%	0.00%
	DIN	Y	43.02%	21.45%
	Strength	N	35.67%	0.00%
	Frequency	N	41.97%	0.00%
	Dose	N	65.18%	0.00%
	Duration	Y	59.65%	38.20%
	Dispensed Count	Y	91.87%	47.45%
	Dispensed Form	Y	55.58%	47.45%
	Refill Count	Y	75.93%	66.98%
	Start Date	–	99.67%	
	Stop Date	–	58.83%	
	Date Created	–	99.93%	
MEDICAL PROCEDURE	Name	N	100.00%	0.00%
	Code	N	0.00%	0.00%
	Performed Date	–	80.11%	
	Date Created	–	96.31%	
LAB	Name	Y/N	99.95%	68.64%
	Code	Y/N	75.72%	68.64%
	Test Result	Y/N	93.55%	53.79%
	Performed Date	–	95.83%	
	Date Created	–	100.00%	
PATIENT	Sex	Y	99.96%	99.87%
	Birthyear	Y	99.89%	
	Status	Y	99.80%	83.68%
	Occupation	N	2.71%	0.00%
	Ethnicity	N	4.58%	0.00%

	FSA /Urban or Rural	Y	89.18%	
	Date Created	–	59.56%	
PROVIDER	Sex	Y	N/A	96.19%
	Birthyear	Y		66.27%
	Provider Type	N		100.00%
	Start Date	–		100.00%
RISK FACTOR	Name	Y	100.00%	88.19%
	Status	Y	25.60%	64.97%
	Frequency	N	3.23%	0.00%
	Duration	N	0.01%	0.00%
	End Duration	N	0.00%	0.00%
	Start Date	–	21.43%	
	End Date	–	1.98%	
	Date Created	–	64.39%	
REFERRAL	Name	Y	100.00%	60.74%
	Concept Code	Y	100.00%	60.74%
	Completed Date	–	24.43%	
	Date Created	–	100.00%	
VACCINE	Name	Y	99.72%	90.29%
	ATC Code	Y	47.48%	90.29%
	DIN	Y	18.91%	15.56%
	Dose	N	37.76%	0.00%
	Not Given	Y	100.00%	100.00%
	Not Give Reason	N	0.08%	0.00%
	Reaction	N	9.01%	0.00%
	Admin Site	N	43.57%	0.00%
	Route	N	28.13%	0.00%
	Lot	N	49.89%	0.00%
	Given Date	–	99.94%	
	Expiry Date	–	25.72%	
	Date Created	–	99.91%	

CPCSSN has developed advanced tools that clean and code the original data extracted from the EMR into a useable format. A detailed evaluation of Table 2 reveals that many of the common and essential data elements (name, code and date) needed for most epidemiologic and clinical studies are being well captured in a coded variable. For these main elements we observe a small decrease in the % present when the data element is converted from an uncoded to a coded field. This modest drop in % present is likely due to entries in the original text or code that the coding tools were not able to standardize. Table 2 also reveals that there is a significant opportunity to expand the completeness of the CPCSSN database through creation of

additional coding and cleaning algorithms. This is particularly the case for variables that have a high % present in the uncoded data but a low % present when coded.

In respect to patient demographic data elements in CPCSSN, the key variables birth year and sex are coded, and more than 99% present. In contrast, the variables containing information on education, occupation, and ethnicity are both largely missing and uncoded. This indicates that these data elements are either not often recorded or are stored in free-text fields within the EMR that is not extracted by CPCSSN. Presently, there is little opportunity to increase the completeness of this type of data without a change in documentation patterns by primary care providers, or the use of advanced NLP and ML techniques to pull this information from clinical notes.

The raw diagnostic data (diagnosis text and/or diagnosis codes from the EMR) reveals that information about a diagnosis associated with a clinical encounter (as found in the Encounter Diagnosis or Billing tables), or diagnostic data in a patient's problem list (a list of a patient's active diagnoses and key health issues) is generally present (<6% null for the uncoded diagnosis fields). For the coded versions of these fields, there is a lower % present (75% present for the Encounter Diagnosis and Billing tables, and 45% present for the Health Conditions table). This highlights an opportunity to improve the diagnoses coding, particularly for the Health Conditions table, which lists a patient's key health issues.

Raw EMR medication and lab data is well captured within the database (very high % present in the uncoded data for names, dates and other variables), as are the coded medication names and ATC code fields. However, there is still considerable work to be done to standardize information from the original EMR data fields. Specifically, CPCSSN only standardizes data from 53 labs, thus there is an opportunity to expand the list of labs that are coded and cleaned. Additionally, CPCSSN coding could be extended to include medication metrics such as frequency, dose, and duration.

Beyond the original text name of the data under consideration, the quality of data on allergy intolerances, medical procedures, vaccines, and some risk factor specifications (frequency and duration) within the CPCSSN database is poor, as is evident by the low % present in even the uncoded fields. Much of this data is often undocumented or input into an unstructured field within the EMR, and thus not easily extracted and pulled into the CPCSSN database. In the case of medical procedure data, while CPCSSN does extract some of these data, presently there is no process to code it to a standard ontology.

It is essential to account for lifestyle, risk factors and socioeconomic status (SES) in clinical and epidemiological research, and in the evaluation of policies, programs and services designed to improve the health of individuals or populations.[26] Research is revealing the importance of these factors as key determinants of health.[26] Unfortunately, detailed information about individual level SES measures such as occupation or income, and lifestyle factors such as physical activity level, smoking, and substance use, are not always documented in the EMR. If

they are documented it can often be found in free-text domains within the EMR, such as in the SOAP (*Subjective, Objective, Assessment and Plan*) notes. SOAP notes are not extracted by most CPCSSN networks as expanded research ethics approvals would be required. For some risk factors that are linked to a patient's geographical location, there are area-based-economic indicators that can be utilized to track effects of SES on health.[26]

Table 3 and Figure 2 display the completeness of several key covariates within the CPCSSN database for a group of active patients (recent visit to primary care; see methods sections for denominator definition). Due to privacy constraints and varying provincial legislation, CPCSSN does not receive sufficient information from all PBRLNs that contribute data that would enable the derivation of an area-based SES measure. Promisingly, CPCSSN is currently developing a process that would allow the derivation of an area-based SES measure (Pampolon, Canadian Marginalization Index, Canadian Index of Multiple Deprivation) without obtaining and storing the full postal code so that this covariate can be available for a higher proportion of patients within the database. Most PBLRNs do submit data on the rural or urban location of a patient, which is derived from the second digit of the postal code (0 indicates rural, all other numbers indicate urban).

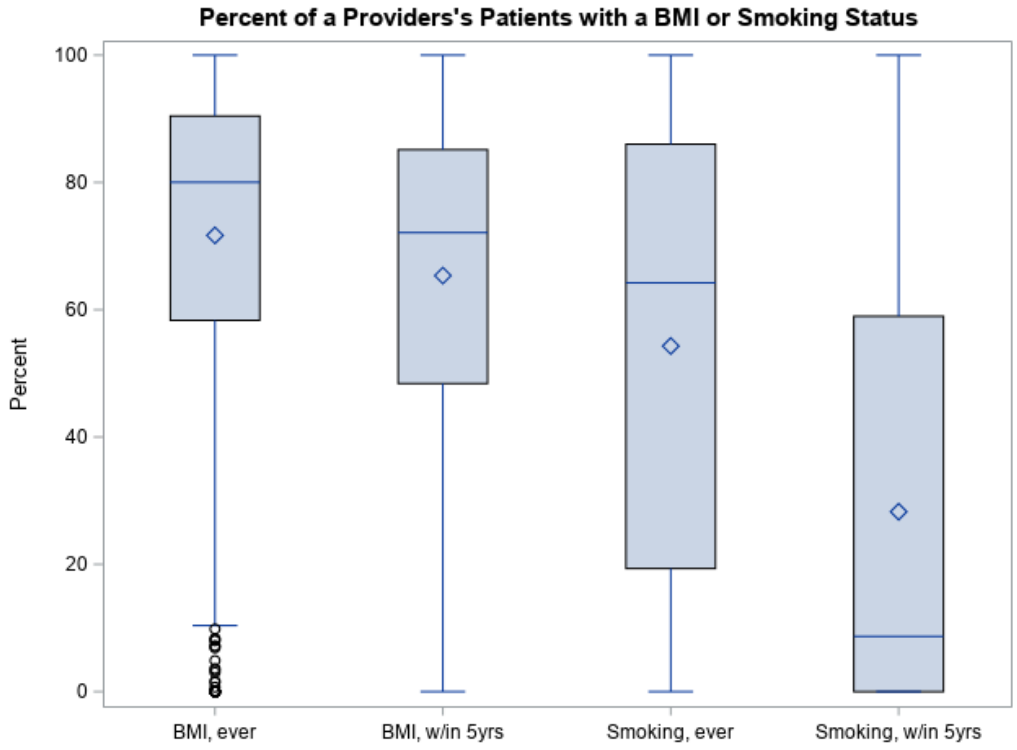
There is a significant proportion of patients missing a documented BMI (29%) or smoking status (53%). Figure 2 reveals broad variation in documentation patterns of BMI and smoking status between providers.

Some of the limitations of the missing data elements can be compensated for using evidence-based approaches aimed at understanding and classifying the mechanisms underlying the missingness (missing-at-random, not-missing-at-random), which can then help drive effective analytical adjustment methods.[27,28]

Table 3. Presence of Geographically Derived Measures

Covariate	% PRESENT
Social and Material Deprivation Score (quintiles)	47.15%
Rural vs. Urban Location	93.42%

Figure 2. Practice Level Documentation of BMI and Smoking



(B) Element Agreement

Description: Data element agreement refers to a comparison of two or more elements within an EMR to determine if they report the same or compatible information. This is a measure of accuracy and can be assessed using data quality probes (DPQs), where a question is asked of the data to highlight disagreement between parts of the patient record (e.g., how many children have an Alzheimer's diagnosis).[13,14] In a database with excellent data quality, we would expect DQPs to return a very low number of records (<1%).

Calculation description: The DQP will evaluate both original and calculated fields within the database.

Type of measure: Percentages.

Findings

There are various clinical and social demographic data elements within the EMR that describe a patient and their health. These data elements may be incongruent due to an error in the patient encounter system (from provision of healthcare and data entry into the EMR, to data retrieval and processing by CPCSSN). Five DQPs were applied to an active patient population of 1,199,564 patients within the CPCSSN database (see methods for denominator definition) to evaluate specific cases of incongruencies in the data. These are shown in Table 4.

Table 4. Data Quality Probes

Data Quality Probe		n	Total	Percent
1	Patient is male and has a prescription for birth control	35	539,507	0.01%
2	Patient is male and has a menopause diagnosis	156	539,507	0.03%
3	Patient <18 years old has dementia diagnosis	305	48,732	0.63%
4	Patient has healthcare (exam, billing, prescription, lab) dates < birthdate	161	1,817,213	0.01%
5	Patient has healthcare record (billing, diagnosis, medication, lab, exam) >1 year after deceased date	1,554	20,214	7.69%

Encouragingly, the first four DQPs reveal that there are few instances of prescription or diagnostic incongruencies with age and sex for these specific probes. In addition, there are also few cases of records dated prior to a patient's birth, indicating good date concordance. Surprisingly, there are a higher-than-expected number of records dated one year or more after a patient's death. This discordance could occur at several points in the encounter system, including a delay in a primary care provider being notified about a patient's death, data entry error (transposition of numbers) or data conversion errors when a clinic migrates from one EMR system to another (this can result in resetting of record dates). Data conversion errors are not uncommon and may explain the high number of records identified for this DQP.

(C) Data Source Agreement

Description: This indicator evaluates how information derived from CPCSSN compares to other sources of information.[4]

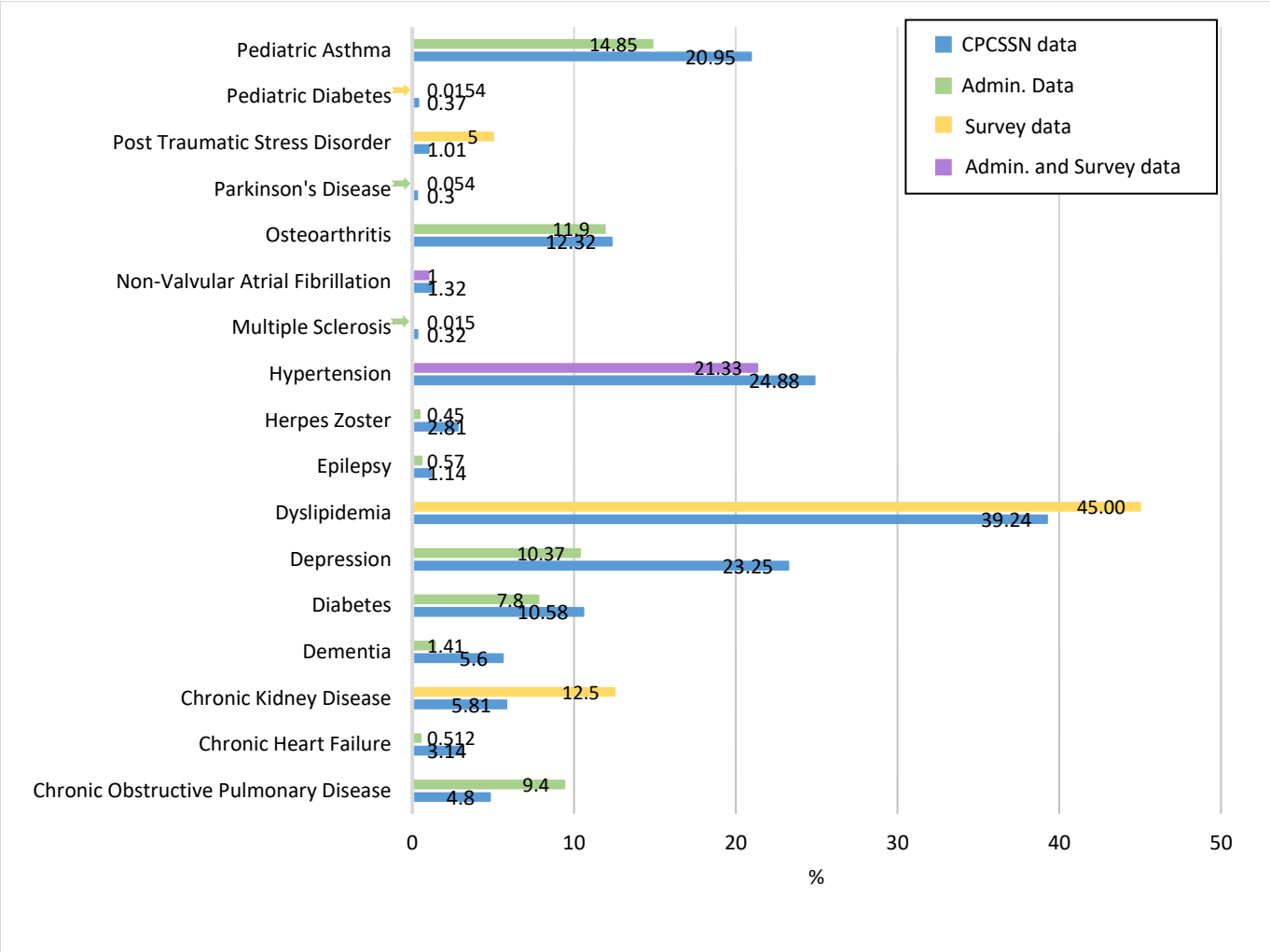
Calculation description: A comparison of the prevalence of some common chronic diseases estimated using CPCSSN data (age and sex adjusted to the 2016 Canadian Census), to estimates from other sources (administrative data, survey data, administrative and survey data).[29-45]

Type of measure: Proportions, with variance estimates.

Findings

Researchers working with CPCSSN data have developed and validated many case detection algorithms to identify and classify patients with common chronic diseases.[15-22] There are currently 18 case definitions that are implemented and available within the database.[15-22] Figure 2 displays the age and sex adjusted prevalence of these 18 conditions as determined from the CPCSSN database, compared to estimates of these conditions from other sources.[15-22,29-45] It can be difficult to compare disease prevalence estimates that are derived from different data sources, as even following age and sex adjustment, other data, methodological, or definitional differences cannot always be compensated for.

Figure 2. Health Condition Prevalence



The disease estimates derived from the CPCSSN data are reasonable and comparable to estimates from other sources.[29-45] Some condition estimates more closely align with estimates from other sources (e.g., diabetes, hypertension, osteoarthritis), while others show more variation (e.g., depression, PTSD).[29-45] It is well known that some conditions are inherently difficult to measure and evaluate, and this is evident in the difference between the CPCSSN estimates to the comparable sources. Conditions with less objective and measurable parameters, like mental health conditions, show increased variability in their prevalence estimates. In contrast, conditions such as diabetes can be more reliably and objectively diagnosed and explain why we see a closer alignment of the CPCSSN estimate to the comparable source. Furthermore, the CPCSSN estimates may be closer to the real prevalence in comparison to survey data estimates, which has known biases.

(D) Distribution Comparison

Description: This indicator evaluates distributions or summary statistics of aggregated data from the EMR and compares them with expected distributions for clinical concepts of interest. [4] This indicator will evaluate the distribution of exam results.

Calculation description: Measures of central tendency (i.e., mean, median, interquartile range, standard deviation) for the coded exam measures within the database: BMI, Waist Circumference, and Systolic and Diastolic Blood Pressure**.**

Type of measure: Statistics.

Findings

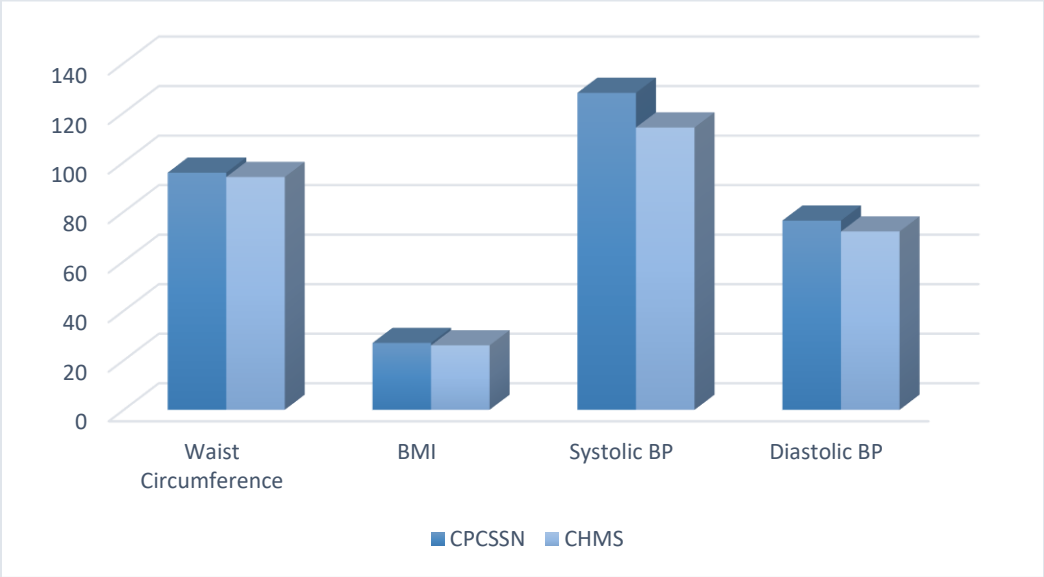
The distribution comparison indicator assesses the distribution of specific clinical measures to determine if the data are within clinically plausible parameters. Evaluating waist circumference observations, BMI observations, and blood pressure observations recorded between 2010 and 2022 reveals the values within the database fall within expected parameters (see Table 5). A comparison of these exam measure to those reported by the Canadian Health Measures Survey (CHMS) is displayed in Figure 3.[45,46] The CMHS 2012-2013 reports the average waist circumference of females at 90.5 cm and males at 97.5 cm, which is very close to the central measure of waist circumference observations in the CPCSSN database.[45] The distribution of BMI observed in the CPCSSN data (50% of patients have a BMI > 26) is comparable to the proportions reported by the CHMS (62% of Canadians have a BMI >25).[46] There is some incongruence with the blood pressure distribution when compared to the CHMS estimates.[45] The distributions observed in the CPCSSN database reveals that the average systolic and diastolic (systolic, mean=128mmHG; diastolic, mean=76mmHG) is higher than reported by the CHMS (systolic, mean=114mmHG; diastolic, mean=72mmHG).[45] This may be a result of sampling bias as patients with hypertension (high blood pressure) are more likely to have a blood pressure measurement in the EMR than those with normal or low blood pressure.

Table 5. Distribution of Exam Observations

Exam	*n*	*Mean*	*SD*	*Median*	*IQR*	*Q1*	*Q3*	*Min*	*Max*
Waist Circumference (cm)	290,456	95.76	19.25	96	22	85	107	30	300
BMI (kg/m^2)	5,379,161	27	9	26	11	21	31	10	100
Systolic BP (mmHg)	9,453,470	128	18	127	23	116	139	50	300
Diastolic BP (mmHg)	9,502,067	76.4	11.01	77	13	70	83	30	200

BMI=Body Mass Index, BP=Blood Pressure, n=number of observations, SD=Standard Deviation, IQR=Interquartile Range

Figure 3. Comparison of CPCSSN Exam Measure to Canadian Health Measures Survey (CHMS)



(E) VALIDITY CHECK

Description: This indicator assesses the plausibility of the values within the EMR.[4] To do this we evaluated the with-in person variation for two clinical exam measures.

Calculation description: a) in patients with at least two BMI measurements in the last five years, evaluate the change in BMI between two independent measures; b) in patients with at least two BP measurements in last five years, evaluate the change in BP between two independent measures.

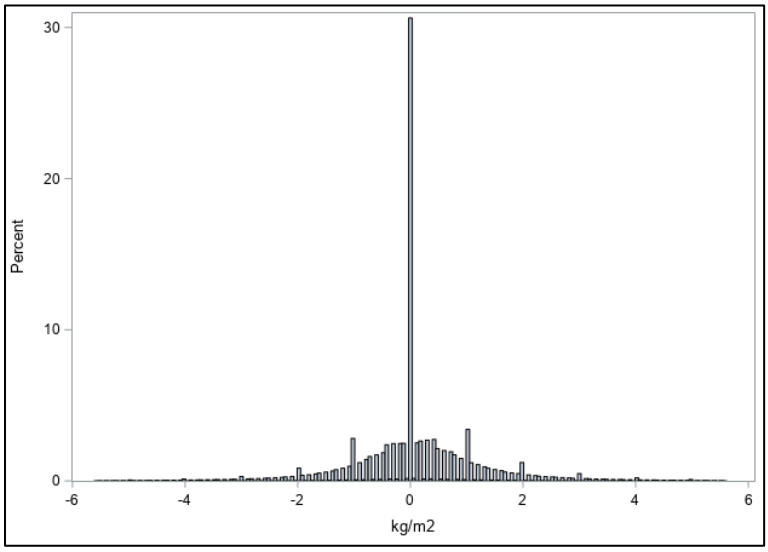
Type of measure: Distribution.

Findings

While much of the data within the CPCSSN database may be within clinically plausible parameters, that does not mean these data are all accurate. One way to assess the accuracy, hence validity, of the data is to evaluate measures of 'within-person' data. The examples we focus on here is the with-person variation between consecutive BMI and blood pressure observations.

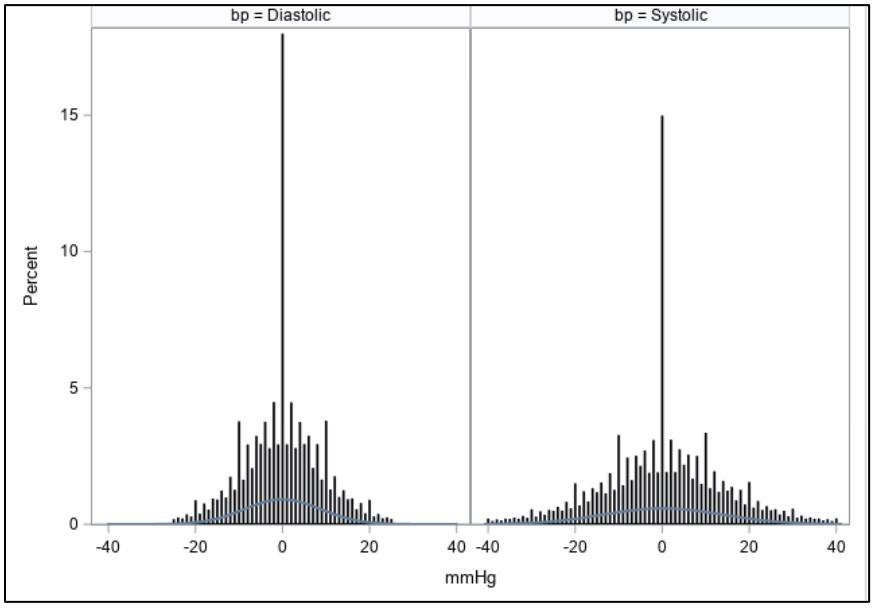
The overall distribution of the within-person changes in BMI between one observation and the next can be seen in Figure 4. The data reveal that 99% of the changes in BMI from one observation to the next fall within a range of [-5.6 to +5.6], only these top 99% are shown in Figure 3. This distribution provides confidence that the vast majority of BMI observations are consistent, as almost all measures fall within two BMI units of a previous BMI observations. Many of the BMIs changes outside the top 99% may still be plausible, as they could occur in patients who have had bariatric surgery (procedure performed on stomach or intestines to induce weight loss); however, it may also be a result of data or processing errors.

Figure 4. Within-person change in BMI between consecutive measurements



The overall distribution of the within-person changes in blood pressure from one observation to the next can be seen in Figure 5 (extreme observations [>25 or <-25 mmHg, for diastolic], [>40 or <-40 mmHg, for systolic] were removed). The data reveal that 98% of observations are within 25 mm/Hg of their previous diastolic measure, and within 40 mm/Hg of their previous systolic measure. This distribution shows a range that is within expected parameters (systolic measures range from 90 to 140mmHg; diastolic measures range from 60 to 90 mmHg). The 2% of observations that fall outside the plausible parameters are likely data entry, extraction, or unit conversion errors. These results provide support for the validity of blood pressure measures in the CPCSSN database.

Figure 5. Within-person change in blood pressure between consecutive measurements



3. COMPARABILITY AND COHERENCE

This data quality dimension describes the degree to which information is comparable over time and across jurisdictions and how easily it can be combined with other sources.[7,9] This dimension is of particular importance as CPCSSN is a repository of data from many jurisdictions (seven provinces), extracted from ten different EMR products. This quality dimension will be evaluated using a single indicator: Data across Sites, Provinces and EMR Type.

DATA ACROSS SITES, PROVINCES AND EMR TYPE

Description: This indicator will provide information on the degree to which information from each network (province, EMR and site) is comparable across sites, province, and EMR type.

Calculation description: Percent of patients with common diagnostic codes; percent of patients with common medications; percent of patients with common labs.

Type of measure: Proportions.

Findings

To understand the comparability and coherence of the data, we have created an indicator that compares the percent of patients active in 2018-2019 (see methods section for denominator definition) with common diagnostic codes (found in the Billing, Encounter Diagnosis or Health Conditions tables), medications, and labs by (1) site and province; and (2) site and EMR type. These indicators are a particularly important component of this data quality report as they can highlight important sources of heterogeneity in the CPCSSN database. It should be noted that province and EMR type are not independent, as some EMR types are only used in certain provinces.

To construct the indicator, the 12 most common diagnostic codes (ICD-9), medications (ATC codes) and labs (LOINCs), as found in the coded fields, were selected from the CPCSSN database. These are listed in Table 6.

Table 6. Most common diagnoses, medications, and labs in the CPCSSN database

Diagnoses (ICD-9)		Medications (ATC)		Labs (LOINC)	
401	Essential Hypertension	N06	Antidepressants	718-7	Hemoglobin (Hb)
250	Diabetes Mellitus	N02	Analgesics	777-3	Platelets in Blood (PLT)
V70	General Medical Exam	J01	Antibiotics	26464-8	Leukocytes
300	Anxiety	C09	ACE Inhibitors	788-0	Erythrocyte Distribution Width [Ratio] (RDW_CV)
311	Depressive Disorder	C10	Lipid Modifying Agents	789-8	Erythrocytes in Blood (RBC)
V20	Health Supervision of infant or child	A02	PPI	14682-9	Creatinine (Serum-Cr)
799	Other ill-defined and unknown cause of morbidity or mortality	R03	Adrenergic	787-2	Erythrocyte Mean Corpuscular Volume (MCV)
780	General Symptoms	N05	Psycholeptics (antipsychotics, anxiolytics, sedatives)	786-4	Erythrocyte Mean Corpuscular Hemoglobin Concentration (MCHC)
460	Acute nasopharyngitis	A10	Diabetic Drugs	12195-4	Plasma Creatinine Clearance (eGFR)

V06	Need for prophylactic vaccination	G03	Contraceptives	785-6	Erythrocyte Mean Corpuscular Hemoglobin (MCH)
781	Symptoms involving nervous or skeletal system	M01	Anti-inflammatory and antirheumatic products	178566-6	Hemoglobin A1c (HbA1c)
650	Normal Delivery	N03	Antiepileptics	1742-6	Alanine Aminotransferase in Serum or Plasma (ALT)

Diagnostic Codes (ICD-9)

Each graph in Figure 6 is a comparison of the prevalence of twelve of the most common ICD-9 diagnostic codes used at each CPCSSN site, by province (6a), and by EMR type (6b).

Comparing the prevalence of these common diagnostic codes shows that there is significant variation in coding practices across sites, and these differences are influenced by province and/or EMR type. The graphs reveal that there is wide variation in the use of more general codes (e.g., V70, General Medical Exam; 780, General Symptoms), with less variation for more disease specific codes (e.g., 250, Diabetes Mellitus; 311, Depressive Disorder). This is not true across the board, as some more condition-specific codes, such as Acute Nasopharyngitis (460), still show large variation in use by site, EMR type and province.

This variability by site, province and EMR, in addition to the different trends between types of ICD-9 diagnostic codes (i.e., general, and specific), reveals that the variation in diagnostic code prevalence arises from multiple mechanisms. The data indicate that these mechanisms likely include differences in data discipline (i.e., documentation patterns) at the site level, but that these differences may be influenced by province (e.g., billing practices/provincial or territorial billing requirements, clinic resources, clinical guidelines) and EMR type (i.e., ease of use, coding flexibility). This means that the diagnostic codes may be used inconsistently, and users of the pan-Canadian database should use caution when relying on a single ICD-9 code to group patients within researcher-defined diagnostic classes. While these results suggest further work is needed to clean and code the diagnostic data within the pan-Canadian database, there is also considerable evidence to suggest that much of the inconsistency is due to data discipline and provider coding practices, themselves.

Figure 6a. Prevalence of the 12 most common diagnostic codes recorded at each site, by province

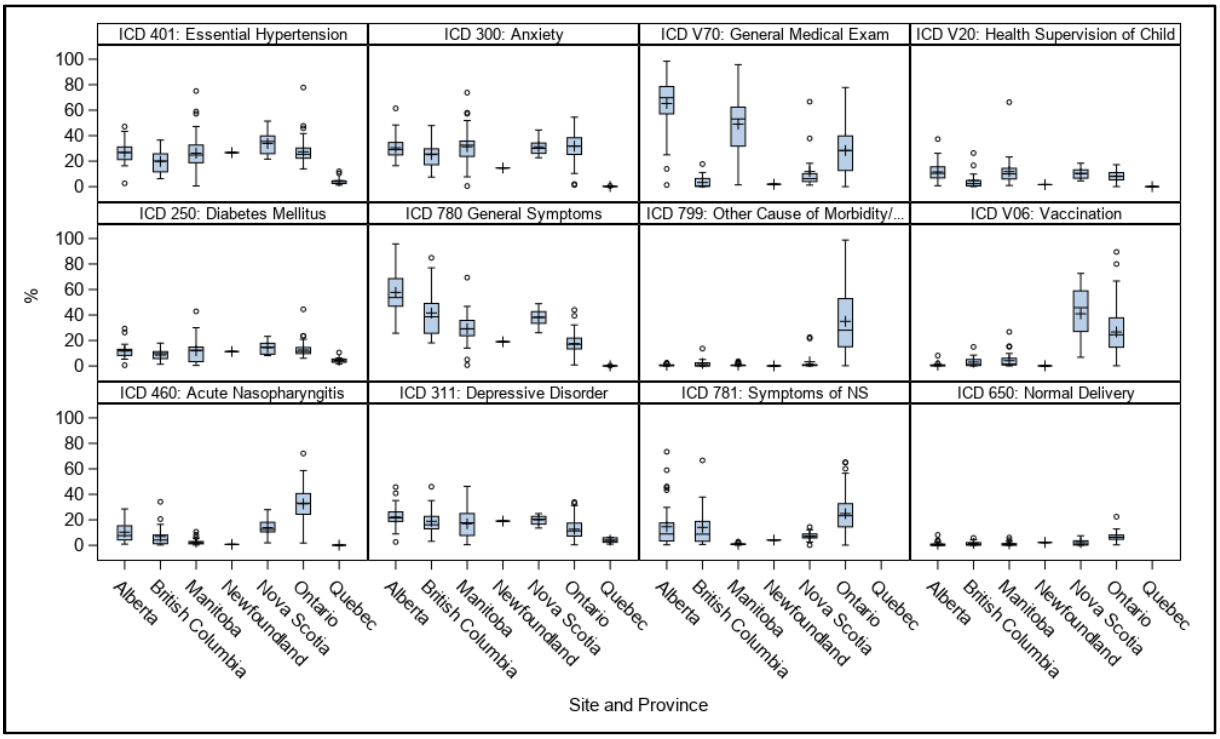
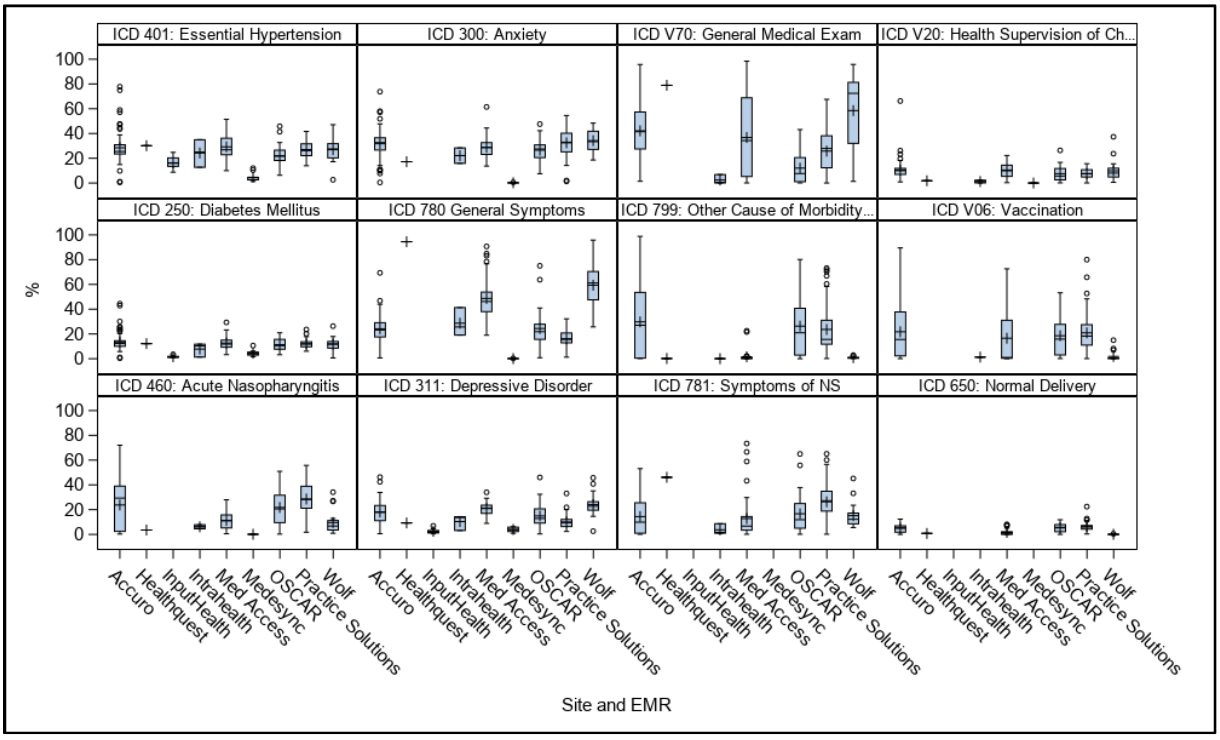


Figure 6b. Prevalence of the 12 most common diagnostic codes recorded at each site, by EMR type



Medication Codes (ATC)

Figure 7 compares the prevalence of the twelve most common medication classes prescribed at each site, by province (7a) and by EMR type (7b). There is a low amount of variation in the prescribing and documentation of medications between sites, province and EMR type, as compared to the ICD-9 coding. However, comparing different medication classes reveals that there is wider variation in some classes of medication, particularly antibiotics (J01), anti-inflammatory drugs (M01) and analgesics (N02).

The above findings suggest that the prescribing data within the CPCSSN database is of good quality as there is little evidence to suggest differences in documentation patterns. However, further research is required to determine if the observed variation reflects true differences in prescribing patterns across sites, which may be influenced by clinic culture, and by provincial guidelines and programs.

Figure 7a. Prevalence of the 12 most common medication codes prescribed at each site, by province

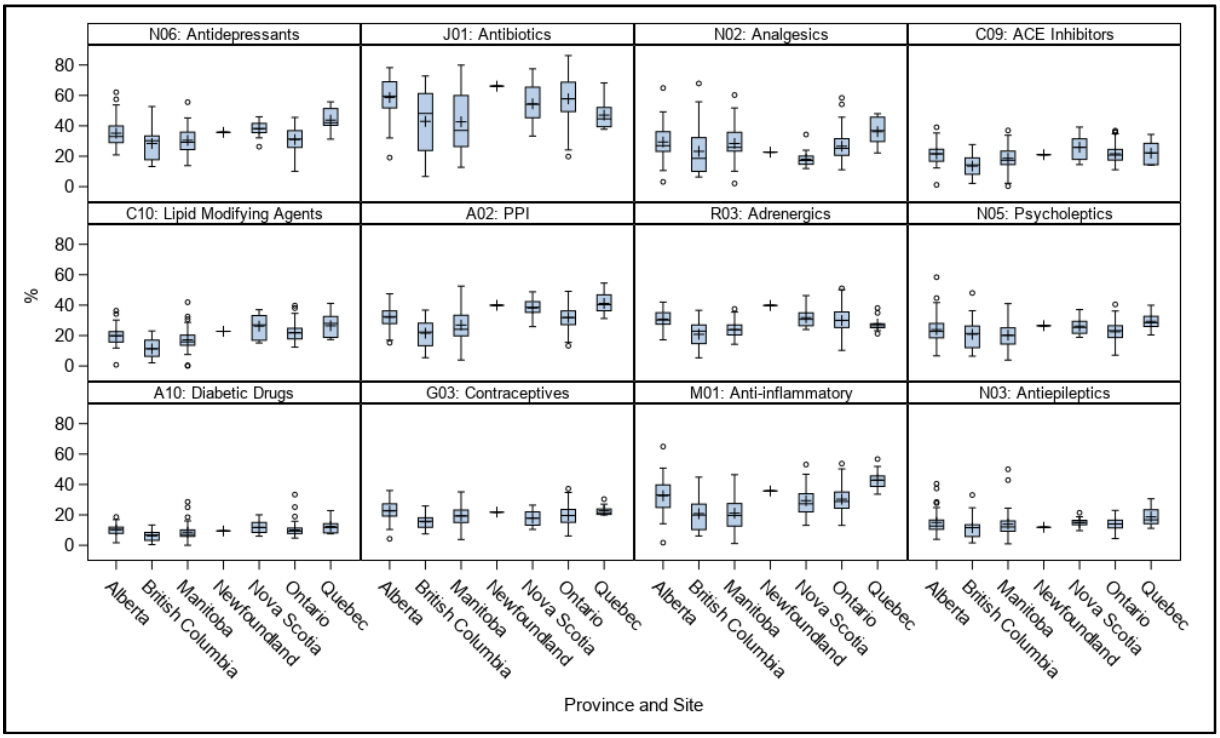
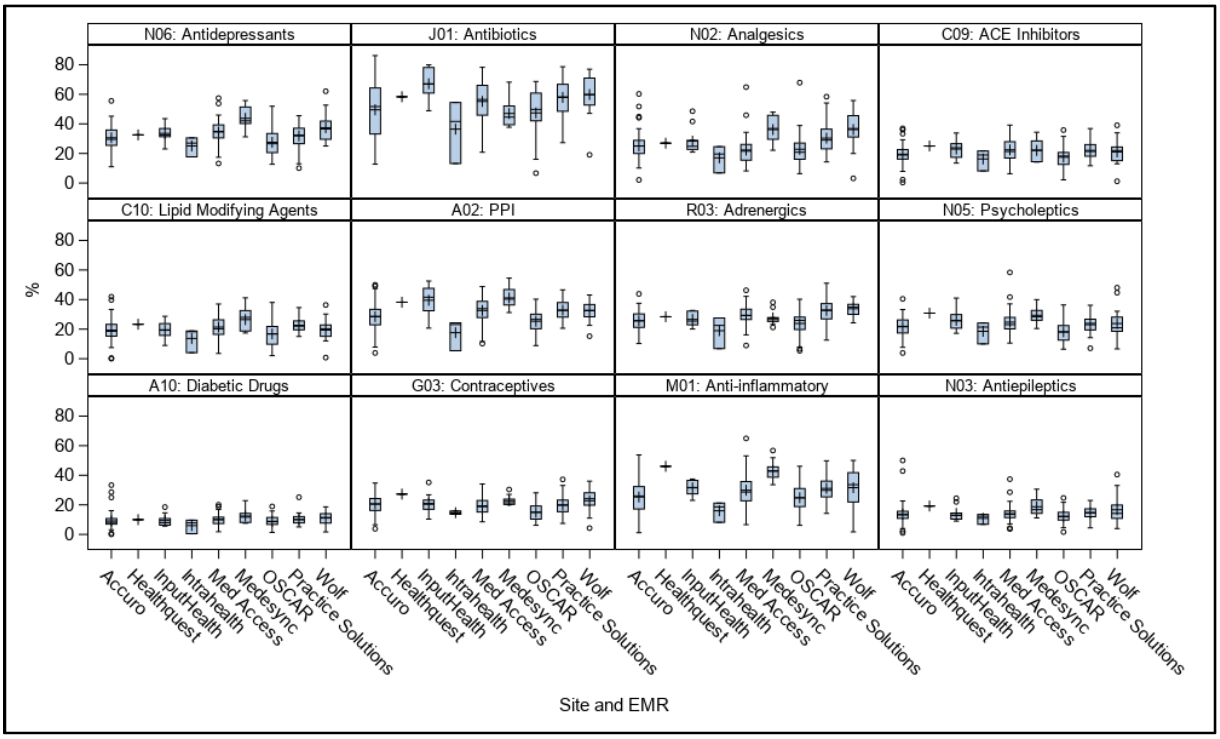


Figure 7b. Prevalence of the 12 most common medication codes prescribed at each site, by EMR type



Lab Codes (LOINC)

Figure 8 shows a comparison of the prevalence of the twelve most common labs recorded at each site, by province (8a), and by EMR type (8b).

Comparing lab prevalence across sites, by EMR type and province, reveals stark variation. The broad differences by site, and across EMR type and province, suggests that the quality of the laboratory data is inconsistent. Suspected mechanisms for this inconsistency include differences in frequency of lab requests (potentially due to difference in lab requisition forms), EMR lab data format (pdf and HL7 data is not routinely extracted by CPCSSN) and variability in lab names (not transformed into a standardized format by CPCSSN). As with the diagnostic data, users of the pan-Canadian database should use caution when using lab data as a single information point to determine a patient's diagnosis or to assess quality of care. This comparison indicates that there needs to be further exploration of where the lab data may be stored within the EMR so it can be effectively extracted and incorporated into the CPCSSN database, as well as increased resources devoted to cleaning and coding this data type.

Figure 8a. Prevalence of the 12 most common lab codes recorded at each site, by province

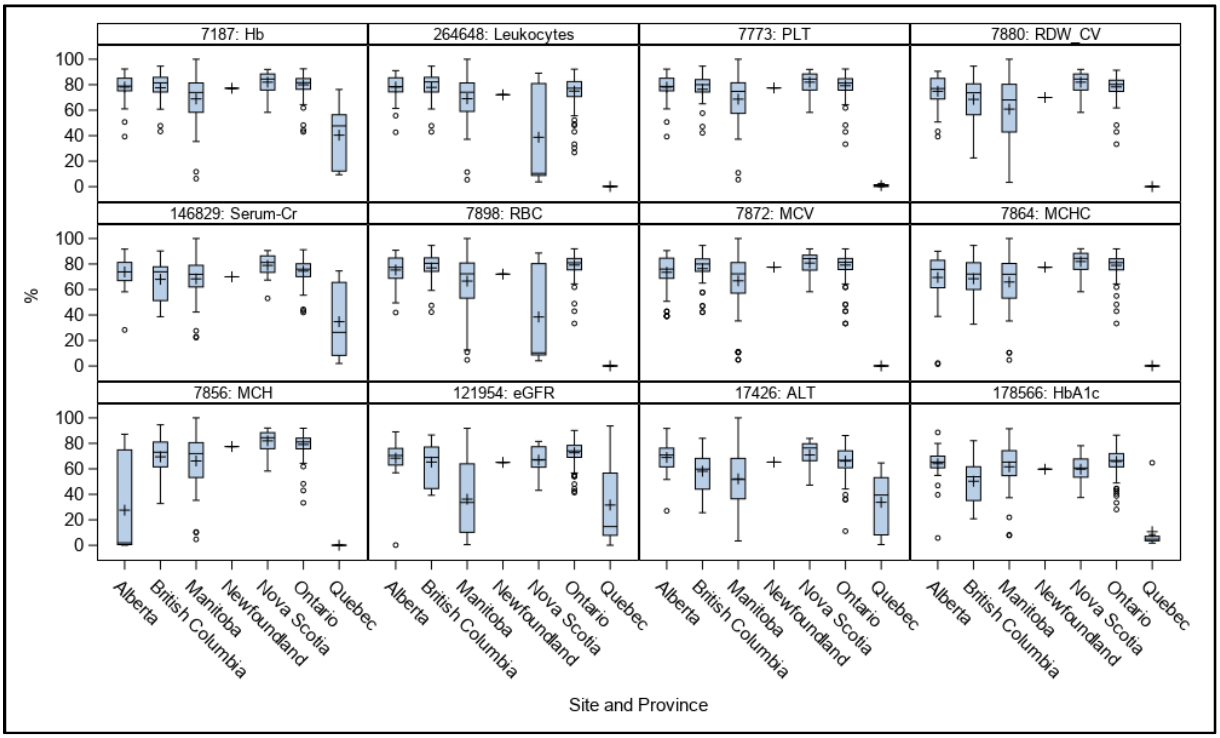
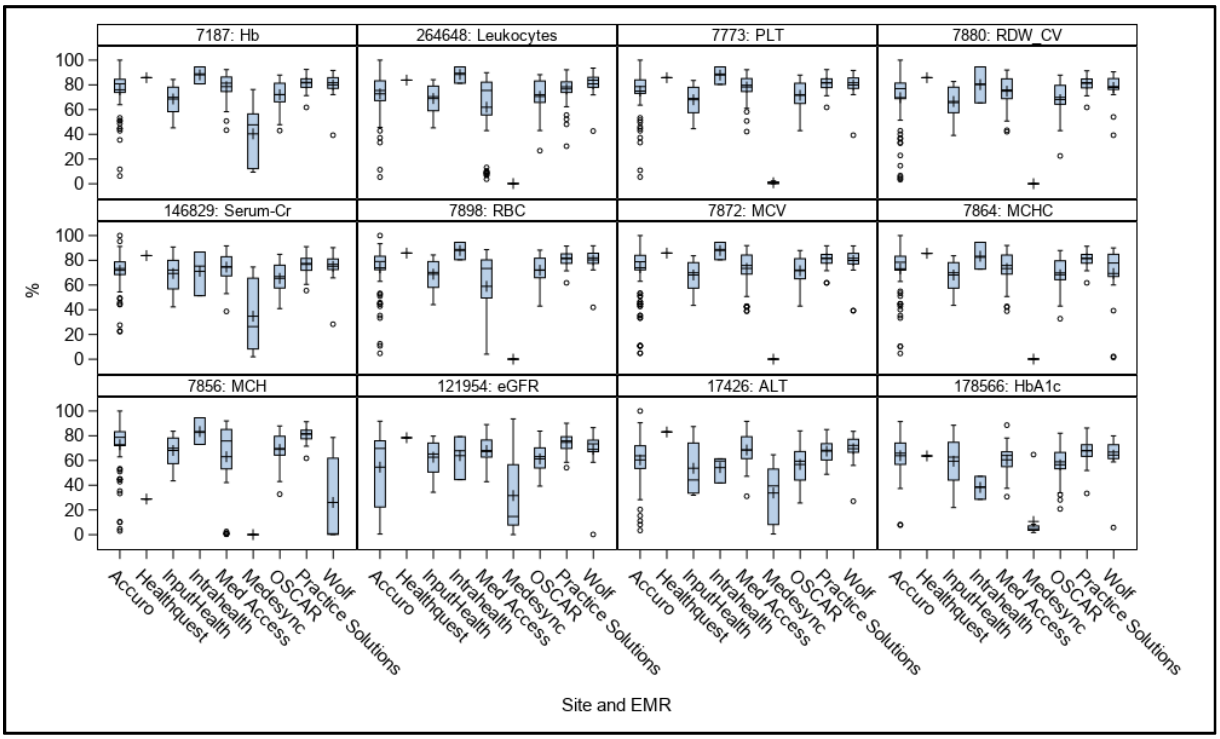


Figure 8b. Prevalence of the twelve most common lab codes recorded at each site, by EMR type



4. TIMELINESS AND PUNCTUALITY

Timeliness refers to how quickly information is made available after the end of each six-month reference period, while punctuality refers to whether information is delivered on the announced dates.[7,9] This quality dimension will be evaluated using one indicator: Data Extraction Frequency and Time to Access.

DATA EXTRACTION FREQUENCY AND TIME TO ACCESS

Description: This indicator provides information on the how often clinical data is extracted from the EMRs of participating practices and the length of time taken to process the data before it is available for researchers.

Calculation description:

Processing: Data are available approximately three months after the extraction date (about April 1 for the December 31 extract; and October 1 for the June 30th extract);

Data Access: How long, on average, the data access process takes, from submission of a data access request form to release of data into the secure research environment (SRE).

Type of measure: Descriptive.

Findings

CPCSSN extracts and processes EMR data on a bi-annual schedule, commencing on January 1 and July 1 of each calendar year, with each data set comprising all historical data up to the extraction commencement date (December 31 for the January 1 and June 30 for the July 1 extractions). Processing is a complex task, taking about 12 weeks to complete, and so the research-ready data set is available by April 1 and October 1 of each year.

Each of the 13 PBLRNs within CPCSSN follows a similar processing pipeline to provide data to a central repository that ultimately becomes CPCSSN's research-ready database. The data are first extracted from the 268 clinic locations that may use one of ten different EMR software products, with the process depending on whether the data are locally hosted by the clinic (i.e., held on-site) or by the EMR vendor. This is the most time-consuming portion of the processing pipeline, typically taking about eight weeks. Locally hosted clinic data may be extracted by CPCSSN data managers (DMs) or by clinic personnel, with the exact procedure and timing varying by clinic configuration and availability. Vendor supplied data generally arrives between two and eight weeks following the commencement date, the timing of which is dictated by the vendor.

Once the data are extracted, each PBLRN then transforms and loads the data into a CPCSSN-style database, typically taking less than a week to complete. Afterward, the individual data sets are cleaned, using a tool that harmonises the raw EMR data between networks, converts it to standard ontologies (e.g., ICD-9, ATC, LOINC), applies CPCSSN case definition algorithms, and deidentifies the data, among many other standardisation procedures. This process also takes approximately one week to complete. Finally, each PBLRN merges the cleaned and standardised data from each site into a single PBLRN-level database. These databases are then

submitted to a central staging repository, where the data are screened for consistency in structure and standardisation. Following the screening and any corrections which must be made, the data are then moved into the main CPCSSN database, where they are ready for use by researchers.

The average time between submission of a data access request and provision of the data via the Secure Research Environment (SRE) is approximately 85 days. However, recently CPCSSN has devoted resources to improving their processes and in 2022 the average time to access CPCSSN data dropped to 44 days. Furthermore, if a data access request is submitted within two weeks of the Data Access Selection Committee (DASC) meeting (monthly) the data can be available in as little as 24 days (assuming approval is granted).

5. ACCESSIBILITY AND CLARITY

This dimension describes the degree to which information, including supplementary or explanatory information and metadata, is easily obtainable and how clearly it is presented.[7,9] The indicator for this dimension will be a description of CPCSSN's Data Stewardship.

DATA STEWARDSHIP

Description: This indicator is a summary of the documentation and practices CPCSSN has developed to ensure the data is accessible, usable, safe, and trusted.

Calculation description: List and description of policies, procedures and user support documents that help researchers access, understand, and use the data.

Type of measure: Descriptive.

Findings

Data stewardship encompasses practices that ensure an organization's data is accessible, usable, safe, and trusted, and this has become an increasing priority for CPCSSN.[47,48] In the early years of CPCSSN, resources were funneled towards the development and creation of the database. Now well into its second decade of existence, CPCSSN is shifting more of its resources to securing and maintaining the integrity of the data within the database while promoting and supporting the use of the data. To that end, CPCSSN has renewed its Governance Framework and has been working on revised policies on security, quality, and access.

Below is a description of current practices, policies, procedures, and tools that support CPCSSN's data governance.

- The **CPCSSN Security Policy** provides requirements and best practices for internal CPCSSN members around data access and transfer, documentation, confidentiality, and processes for data breaches or releases of potentially identifiable information.
- A **Data Access Policy** has been developed to define the process for data access, outline who has permission to access the data, and how the data is made available. This includes ensuring data is only accessed by academic researchers; industry partners are only able to

receive aggregate results. CPCSSN works closely with its users to ensure that the information requested (via a **Data Access Request** form) is available. However, it limits provision of data elements that may present identification risk, such as original or raw data fields, to specific internal data quality improvement projects. The policy also requires that all data be accessed in the CPCSSN Secure Research Environment (SRE), except for internal uses that meet specific criteria and security requirements. Restriction of the data to the SRE ensures that data is held in a secure system from which only vetted and approved aggregate results and other strictly anonymized data can be exported. This closed system ensures increased security by denying access to the internet.

- The **CPCSSN Data Access Selection Committee (DASC)** makes decisions on data access requests for research and surveillance proposals. The DASC is a team of health researchers and interested individuals who provide research expertise and independent review to potential proposals. All research utilizing CPCSSN repository data must be coherent with the CPCSSN purposes and mission and must be approved by relevant Research Ethics Boards. The researcher will be allowed restricted use of the CPCSSN data according to what is agreed upon in the CPCSSN Information and Data Sharing agreement. The research will not be allowed to use the data obtained for other purposes outside that agreement.
- A **CPCSSN Data Dictionary** is updated after each data extraction cycle (bi-annually). It describes the structure of the database, such as the names, definition and attributes of the data elements contained within each table.
- To further support data users working within the SRE, CPCSSN has developed an **CPCSSN SRE Analyst Guide** that provides information to users on how to access CPCSSN data within the SRE. The information contained within this guidebook continues to be developed and improved to meet users' needs.
- Using the data and information within the CPCSSN database often involves identifying patients with specific disease profiles. This can be challenging due to the heterogeneity of the database. For this reason, CPCSSN has developed and validated sophisticated case definitions that classify patients as having specific conditions and diseases. The specifications and validation statistics for each approved case definition are outlined in the **CPCSSN Case Definitions** document. There is also a **Case Definition Standard Operating Procedure** that sets minimum standards and validation guidelines prior to implementing new or revised old case definitions in the pan-Canadian repository. CPCSSN standards for case definition work exists to promote principles of excellence and a rigorous approach to quality improvement, surveillance, and research for communicable and non-communicable diseases. As of Q2 2022, CPCSSN implements case definitions for 18 different conditions, and lists all patients with these conditions in the Disease Case table.
- When using an EMR database such as CPCSSN, researchers often need to capture populations for the condition under study (denominator). An *Internal Quality Improvement*

Report for CPCSSSN was completed in 2021 on **Methods for defining a patient denominator in the national CPCSSN database: recommendations for best practices**. This document identifies and recommends data-informed methods for appropriately defining a patient denominator in the CPCSSN database.

- Issues of generalizability are a primary concern for research using data from population-based clinical information systems. As such, CPCSSN has evaluated and published a report on the **Representativeness of patients and providers in the Canadian Primary Care Sentinel Surveillance Network: a cross-sectional study.** This publication details how representative the data of the patients and primary care providers in the CPCSSN database are when compared to the Canadian population.[25]
- CPCSSN has created and supports a multi-page website that provides information on their organization and how to access the data (**www.cpcssn.ca**).

Overall, CPCSSN has developed good practices and documents to ensure that the data is accessible, usable, trusted, and secure. However, these practices and documents are not always easily accessible to users and CPCSSN needs to improve the clarity of the data access pathway and what supports (information) is available to help users. As CPCSSN continues to evolve and grow, it is important that data stewardship is a keystone to its central operations.

DISCUSSION

Building a repository of EMR data in Canada, which has a complex and geographically varied healthcare system, is challenging. For more than a decade, CPCSSN has been working to develop and standardize primary care data to ensure it is of sufficient quality to be a valuable source for clinicians, researchers, and policy makers. This data quality assessment found that, if used thoughtfully and carefully, the CPCSSN database has sufficient and high enough data quality to be a valuable resource for research and surveillance.

Undertaking a quality assessment of a database that derives its data from a variety of EMRs, each with its own configuration, and spans several provinces is a difficult and complex task. Nevertheless, using a suite of indicators to capture the five data quality dimensions defined under the NQAF framework, this report has found several areas of strength and weakness within the CPCSSN data set.

Below is a summary of the key findings for each data quality dimension, as well as recommendations for improvement.

Relevance

- a wide spectrum of data types (e.g., diagnostic codes, medications, labs, exams) is captured, which provides access to current and past patient health records and information on healthcare delivery.

- the CPCSSN database contains a large representative sampling of clinical records from across the country, with relevance to both regionally and nationally focused research.
- to remain relevant, CPCSSN must continue to work closely with users, clients, and stakeholders.
- the relevancy of the data for users could also be improved by the creation of a robust and operational methodology to link data elements around a patient encounter (visit).

Accuracy and Reliability

- the quality of the data is high in terms of element agreement, validity, distributions of clinical parameters, and comparison to other data sources.
- the element presence (completeness) indicator highlights the extensive work CPCSSN has done to create coded, standardized information.
- we recommend CPCSSN operations continue to develop their cleaning and processing tools to reduce the missingness in coded fields as much as possible.
- higher priority items include expanding the list of labs that are extracted and coded; and improving the coding of medication metrics (e.g., duration, strength).

Comparability and Coherence

- there is a great degree of variation in the use of common ICD9 codes, medications, and labs at each site, within each province, and by EMR type.
- for population-level epidemiological studies it is recommended that users request identification of site, EMR and province so that clustering at these levels can be accounted for in the analysis.
- in some contexts, researchers may want to consider different analytic approaches on data from each EMR and/or province.

Timeliness and Punctuality

- within the numerous and varying constraints of the Canadian primary care context (separate provincial healthcare systems, vast geography, and variation in EMR configurations) the CPCSSN data resource has effective and reasonable timeliness and punctuality.
- significant and long-term increase in resources would need to be in place to increase data extraction frequency.

Accessibility and Clarity

- CPCSSN has developed a library of supplementary and explanatory information to educate and inform users about the database.

- the accessibility and clarity of the CPCSSN data needs improvement by making the supporting information accessible, available, and more clearly understood.
- we recommend the creation of a training module and/or resource guide, which could include a shared repository of code for data preparation, for researchers and analysts to guide them through CPCSSN and its data holdings, from acquisition to analysis.

Conclusion

Overall, the CPCSSN database has reasonable data quality for epidemiological and population-based research. Due to the complexity and heterogeneity of the database, there needs to be increased support and documentation for users to guide them in the application of the database. This is particularly important if the CPCSSN database is being used to feed back information to clinicians for quality improvement.

Some of the data quality issues uncovered in this report can be addressed through appropriate statistical adjustment methods for population-based studies. However, accurately, and reliably using CPCSSN data to identify specific patients with certain disease profiles, or giving practice level statistics back to a provider, can be challenging without careful data analysis support.

We recommend that data quality indicators on the five dimensions be evaluated after each data extraction cycle or after significant changes to the database schema or transformation processes. Data quality also needs to be clearly and consistently communicated to users to guide them in the use of the database for their project or research question. This may involve working with users and providing data quality measures on a study-by-study basis.

CPCSSN holds a wealth of data and information that is integral to transforming Canada's healthcare system into one that is sustainable, accessible and meets the needs of all Canadians. This report provides a foundation for understanding the quality of the data and information held within the CPCSSN repository so it can be used effectively to support Canada's healthcare system transformation.

References

1. Menear M, Blanchette MA, Demers-Payette O, Roy D. A framework for value-creating learning health systems. *Health Research Policy and Systems*. 2019; 17(79).
2. Scott PJ, Dunscombe R, Evans D, Mukherjee M, Wyatt JC. Learning health systems need to bridge the 'two cultures' of clinical informatics and data science. *J Innov Health Inform.* 2018;25(2):126-131.
3. Fox F, Aggarwal VR, Whelton H, Johnson W. A Data Quality Framework for Process Mining of Electronic Health Record Data. *International Conference on Health Informatics (IEEE).* 2018. doi: 101109/ICHI/.2018.00009.
4. 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;20:144-151.
5. Charnock V. Electronic healthcare records and data quality. Health Info Libr J. 2019;36:91–5. https ://doi.org/10.1111/hir.12249 .
6. Liaw, ST, Guo JGN, Ansari S, Jonnagaddala J et al. Quality assessment of real-world data repositories across the data life cycle: a literature review. *Journal of the American Medical Information Association.* 2021; 28(7):1591-1599.
7. United Nations Statistics Division. *Guidelines for the Template for a Generic National Quality Assurance Framework (NQAF)*. 2012.
8. Office for National Statistics. Guidelines for measuring statistical output quality. 2013.
9. CIHI. CIHI's Information Quality Framework. 2017.
10. Terry AL et al. A basic model for assessing primary health care electronic medical record data quality. *BMC Medical Informatics and Decision Making*. 2019; 19:30.
11. Virdee PS, Fuller A, Jacobs M, Holt T, Birks J. Assessing data quality from the Clinical Practice Research Datalink: a methodological approach applied to the full blood count blood test. *J Big Data*. 2020; 7;96. https://doi.org/10.1186/s40537-020-00375-w.
12. Ni K, Chu H, Zeng L, Li N, et al. Barriers and facilitators to data quality of electronic health records used for clinical research in China: a qualitative study. BMJ Open. 2019;9:e029314. https ://doi.org/10.1136/bmjop en-2019-02931 4.
13. Brown PJ, Warmington V. Info-tsunami: surviving the storm with data quality probes. *Informatics in Primary Care*. 2003; 11:229-37.
14. Brown PJB, Warmington V. Data quality probes – exploiting and improving the quality of electronic patient record data and patient care. *International Journal of Medical Informatics*. 2002; 68:91-98.
15. CPCSSN Team, Case Definitions: Canadian Primary Care Sentinel Surveillance Network (CPCSSN), Version 2021-Q4. February 8, 2022. URL: http://cpcssn.ca./research-resouces/cpcssn-case-definitions-v2.

16. Williamson T, Green ME, Birtwhistle R, Kan S, Garies S, Wong ST, Natarajan N, Manca D, Drummond N. Validating the 8 CPCSSN case definitions for chronic disease surveillance in a primary care database of electronic health records. *Ann Fam Med*. 2014;12(4):367-72.

17. 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). doi:10.1016/j.jcjd.2017.10.060

18. Queenan JA, Ehsani-Moghaddam B, Wilton SB, Dorian P, Cox JL, Skanes A, Barber D, Sandhu RK. Detecting Patients With Nonvalvular Atrial Fibrillation and Atrial Flutter in the Canadian Primary Care Sentinel Surveillance Network: First Steps. *CJC Open*. 2021;3(3):367-371. doi:10.1016/j.cjco.2020.10.012

19. 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*. Published online 2019. doi:10.1016/j.jcjd.2019.02.005

20. Vijh R, Wong S, Grandy M, Peterson S, Ezzat A, Gibb AG, Hawkins NM. Identifying health failure in patients with chronic obstructive lung disease through the Canadian Primary Care Sentinel Surveillance Network in British Columbia: a case derivation study. *CMAJ Open*. 2021; 9(2):E376-E383. doi: 10.9778/cmajo.20200183

21. Marrie RA, Kosowan L, Taylor C, Singer A. Identifying people with multiple sclerosis in the Canadian Primary Care Sentinel Surveillance Network. *Multiple Sclerosis Journal – Experimental, Translational and Clinical*. 2019; 1-9. doi: 10.1177/2055217319894360

22. Zafari H, Kosowan L, Zulkernine F, Singer A. Diagnosing post-traumatic stress disorder using electronic medical record data. *Health Information Journal*. 2021;27(4). https://doi.org/10.1177/14604582211053259

23. Greiver M et al. Developing a method to estimate practice denominators for a national Canadian electronic medical record database. *Family Practice*. 2013; 30(3): 347-354.

24. Manitoba Centre for Health Policy Evluations, & Menec V. Defining practice populations for primary care: methods and issues; (2000): Citeseer.

25. Queenan JA, Williamson T, Khan S, Drummond N, Garies S, Morkem R, Birtwhistle R. Representativeness of patients and providers in the Canadian Primary Care Sentinel Surveillance Network: a cross-sectional study. CMAJ Open. 2016 Jan 25;4(1):E28-32. doi: 10.9778/cmajo.20140128. PMID: 27331051; PMCID: PMC4866925.

26. Poirier MJP. Wilson MG. Rapid synthesis: Identifying how area-based socio-economic indicators are measured in Canada. Hamilton, Canada: McMaster Health Forum, 29 March 2019.

27. Rea S, Bailey KR, Pathak J, Haug PJ. Bias in recording of body mass index data in the electronic health record. *AMIA Jt Summits Transl Sci Proc*. 2013; 214-218.

28. Garies S, Cummings M, Quan H, McBrien K, Drummond N, Manca D, Williamson T. Methods to improve the quality of smoking records in a primary care EMR database: exploring multiple imputation and pattern-matching algorithms. *BMC Med Inform Decis Mak*. 2020; 20: 56.

29. Lix LM, Ayles J, Bartholomew S, Cooke CA, Ellison J, Emond V, Hamm NC, Hannah H, Jean S, LeBlanc S, O'Donnell S, Paterson JM, Pelletier C, Phillips KAM, Puchtinger R, Reimer K, Robitaille C, Smith M, Svenson LW, Tu K, VanTil LD, Waits S, Pelletier L. The Canadian Chronic Disease Surveillance System: A model for collaborative surveillance. Int J Popul Data Sci. 2018 Oct 5;3(3):433.

30. Centre for Surveillance and Applied Research, Public Health Agency of Canada. Asthma and Chronic Obstructive Pulmonary Disease (COPD) in Canada, 2018. Report from the Canadian Chronic Disease Surveillance System. 2018. Ottawa, Ontario, Canada.

31. Centre for Surveillance and Applied Research, Public Health Agency of Canada. Osteoporosis and related fractures in Canada, 2020. Report from the Canadian Chronic Disease Surveillance System. 2021. Ottawa, Ontario, Canada.

32. Centre for Surveillance and Applied Research, Public Health Agency of Canada. Heart Disease in Canada, 2018. Report from the Canadian Chronic Disease Surveillance System. 2018. Ottawa, Ontario, Canada.

33. LeBlanc AG, Gao YJ, McRae L, Pelletier C. Twenty years of diabetes surveillance using the Canadian Chronic Disease Surveillance System. *Health Promotion and Chronic Disease Prevention in Canada: Research, Policy and Practice.* 2019;39(11):306-309.

34. Public Health Infobase, Public Health Agency of Canada. Use of health services for mood and anxiety disorders (annual). Canadian Chronic Disease Surveillance System. 2021.

35. Public Health Infobase, Public Health Agency of Canada. Asthma. Canadian Chronic Disease Surveillance System. 2021.

36. Public Health Infobase, Public Health Agency of Canada. Epilepsy. Canadian Chronic Disease Surveillance System. 2021.

37. Public Health Infobase, Public Health Agency of Canada. Multiple Sclerosis. Canadian Chronic Disease Surveillance System. 2021.

38. Public Health Infobase, Public Health Agency of Canada. Parkinsonism, including Parkinson’s Disease. Canadian Chronic Disease Surveillance System. 2021.

39. Amed S, Dean HJ, Panagiotopoulos C, Sellers EA, Hadjiyannakis S, Laubscher TA, Dannenbaum D, Shah BR, Booth GL, Hamilton JK. Type 2 diabetes, medication-induced diabetes, and monogenic diabetes in Canadian children: a prospective national surveillance study. Diabetes Care. 2010 Apr;33(4):786-91. doi: 10.2337/dc09-1013. Epub 2010 Jan 12. PMID: 20067956; PMCID: PMC2845028.

40. Andrade J, Khairy, P, Dobrev D, Nattel S. The Clinical Profile and Pathophysiology of Atrial Fibrillation. Circulation Research 2014; 144(9): 1453-1468.

41. Russell ML, Dover DC, Simmonds KA, Svenson LW. Shingles in Alberta: Before and After publicly funded varicella vaccination. Vaccine. 2014. 32;47: 6319-6324

42. Statistics Canada. Survey on Mental Health and Stressful Events, August to December 2021. *The Daily*. 2022.

43. Arora P, Vasa P, Brenner D, Iglar K, McFarlane P, Morrison H, Badawi A. Prevalence estimates of chronic kidney disease in Canada: results of a nationally representative survey. *CMAJ*. 2013;185(9):E417-E423.

44. Joffres M, Shields M, Tremblay MS, Gorber SC. Dyslipidemia prevalence, treatment, control, and awareness in the Canadian Health Measures Survey. *Can J Public Health*. 2013;104(3);E252-7.

45. Canadian Health Measures Survey (CHMS), 2012 to 2013.

46. Canadian Health Measures Survey (CHMS), Cycle 5 (2016 to 2017), Cycle 6 (2018 to 2019)

47. Peng G, Privette JL, Tilmes C, Bristol S, Maycock T, Bates JJ, Hausman S, Brown O, Kearns EJ. A Conceptual Enterprise Framework for Managing Scientific Data Stewardship. *Data Sci J*. 2018;17(15).

48. Inau ET, Sack J, Waltemath D, Zeleke AA. Initiatives, Concepts, and Implementation Practices of FAIR (Findable, Accessible, Interoperable, and Reusable) Data Principles in Health Data Stewardship Practice: Protocol for a Scoping Review. JMIR Res Protoc 2021;10(2):e22505.