Supplementary File.
Data Quality Assessment Documentation and Reporting Recommendations (as outlined by Khan [25])

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

Database model/structure	The CPCSSN data are stored as a SQL database (either SQL Server or SQLite format) and currently contains 26 tables: ▪ Eight tables house clinic, health care provider, or administrative information (e.g. patient-provider assignment, non-identifiable clinic information, database version information) ▪ Fifteen tables hold both original EMR data and processed (cleaned and coded) data ▪ Two tables (DiseaseCase, DiseaseCaseIndicator) contain a list of patients who have been indexed with one or more of the conditions for which CPCSSN has a validated definition; these definitions use data derived from the 15 data tables ▪ The Deprivation table is empty, but will be populated in the future. Each table is assigned an integer primary key labelled: <table_name>_ID. Most data tables also contain the columns: ▪ Site_ID (an integer assigned to each primary care practice) and Network_ID (an integer assigned to each PBRN) together form a composite primary key in the Site table, and a composite foreign key in most CPCSSN tables ▪ Patient_ID (randomly assigned integer for each patient) is the primary key in Patient table and a foreign key in most other tables ▪ Encounter_ID is the integer primary key for the Encounter table, but a foreign key in several other tables ▪ Provider_ID is the integer primary key for the Provider table and a foreign key for several tables ▪ All columns are strictly typed, enforced either by the database schema (SQL Server) or external software (SQLite) The Entity Relationship Diagram (ERD), which outlines relationships between all CPCSSN tables, can be found on the CPCSSN website: http://cpcssn.ca/research-resources/cpcssn-data-dictionary-and-erd/ For SAPCReN and NAPCReN, extraction, transformation, coding, and cleaning algorithms are programmed in Python.
Data dictionary	The most recent data dictionary is posted on the CPCSSN website: http://cpcssn.ca/research-resources/cpcssn-data-dictionary-and-erd/
Data Processing / Data Provenance	

Data extraction specifications	The general procedure for EMR data extraction in both SAPCReN and NAPCReN is as follows:
	▪ Access the EMR frontend or backend, either through the vendor or by a CPCSSN data manager
	  o **Wolf & PS Suite:** backend extraction completed by Telus Health
	  o **Accuro & Healthquest:** backend extraction conducted by a CPCSSN data manager using a log-in account for both clinic and EMR database
	  o **Med Access:** frontend extraction by a CPCSSN data manager using a log-in account created by the clinic
	▪ Select data tables and columns of interest, then:
	  o Null any columns with identifiable information (e.g. names, addresses, health care numbers)
	  o Select patients with an assigned CPCSSN provider or for those who have not been assigned any provider, use the four-cut method [31] to assign a provider
	  o An EMR mapping file is generated which consists of four data items: EMR patient ID (number assigned by each clinic EMR system), Network ID (number that uniquely identifies each CPCSSN regional network), Site ID (number that uniquely identifies each clinic within a network) and CPCSSN patient ID (unique patient number randomly assigned by CPCSSN). This file is used to ensure CPCSSN patient IDs are assigned uniquely to each EMR patient ID within each clinic and is held separately from the patient health data. The file can also be used to assist with re-identification for practice quality improvement or linkage with other data sources
	▪ Transfer data to CPCSSN computers via a secure transfer method (e.g. SFTP)

	See Figure 2 for CPCSSN data pipeline.
Mappings from original values to standardized values	The CPCSSN database includes both original data extracted from the EMR, and cleaned or coded information resulting from CPCSSN's algorithms (see CPCSSN Data Dictionary [28]). All original data are extracted from the EMR unchanged, except for a small number of EMR-specific code conversions and some parsing. <u>General cleaning steps</u> ▪ Null empty strings and remove leading or trailing whitespace from all records ▪ Normalize postal codes to format A1A 1A1 o NAPCReN: truncate to forward sortation area code (A1A) ▪ Convert all dates to format yyyy-mm-dd ▪ Verify code formats for ATC, ICD-9, and DIN classification systems ▪ Delete invalid labs where o lab name is a purely numeric value, or matches a text patterns that indicates it is not a useful lab name (e.g. 'unit', 'patient name', 'date') o lab value non-numeric; exceptions are lab values starting with a comparison operator and Hepatitis C ▪ Convert medication dispensed forms and duration units to CPCSSN standard forms ▪ Delete empty or duplicate records ▪ Remove orphaned records (e.g. records with no matching Patient_ID in the Patient table) ▪ Remove patients who only have demographic information and no other data ▪ Parse billing records that contain multiple conditions into individual records <u>Allowable ranges of values for NAPCReN and SAPCReN:</u> ▪ SBP: 50 to 300 mmHg ▪ DBP: 20 to 200 mmHg ▪ Weight: 1 kg to 500 kg ▪ Height: 30 cm to 230 cm ▪ BMI (adult): 5 to 200 kg/m^2 ▪ Waist circumference: 10 to 300 cm ▪ Waist-to-hip ratio: 0.1 to 1 ▪ PEFR: 0-200 L/min ▪ Ranges for laboratory values are specific to each of the 44 lab results processed by CPCSSN ▪ Age at onset: ≤120 ▪ Birth year: extraction year (e.g. 2018) minus 120 years ▪ Record creation date: earliest EMR date (1990-01-01) to extraction end date (2018-12-31) ▪ Condition onset or procedure performed date: 1898-12-31 (extraction end date minus 120 years) to 2018-12-31 <u>Coding to reference system</u> ▪ **Medication**: existing code, DIN, or free text mapped to codes in the ATC Classification System ▪ **Laboratory values:** existing code or free text mapped to LOINC; free-text mapping is completed for 44 labs ▪ **Diagnostic codes:** ICD-9 or other codes (e.g. ICD-10, ICPC) found in Health Condition/Profile, Encounter Diagnosis, and Billing tables mapped to ICD-9 codes/classification heading ▪ **Diagnostic text:** Pattern matches are used to map free text to ICD-9 codes; this mapping can be completed for 48 conditions in the Health Condition/Profile, Encounter Diagnosis, Billing tables ▪ **Referrals**: original text in Referral field mapped to SNOMED concept code; not extracted from referral letters or any PDF documents ▪ **Risk factors:** information in the Risk Factor table categorized by text pattern matching to one of the following headings: Smoking, Alcohol, Exercise, Diet, Obesity, or Psychosocial Stress.

	Smoking and Alcohol information further parsed (when available) into Status, Dose, Frequency, Duration, End date ▪ **Family history:** relationship type is identified by pattern matching as one of: Mother, Father, Daughter, Son, (Half) Sister, (Half) Brother, (Great) Grandmother, (Great) Grandfather, (Great) Aunt, (Great) Uncle, Niece, Nephew, Cousin, or None. Relationship side (Maternal or Paternal) is also identified by text pattern matching
Data management organization's data transformation routines, including constructed variables	CPCSSN Data Dictionary [28] highlights CPCSSN-constructed variables in blue, including: ▪ **Calculated age:** current year minus year of birth *(only as part of the national data provided to approved data users)* ▪ **BMI**: uses original BMI if available; otherwise, CPCSSN algorithms generate BMI from original height and weight data where none previously exists ▪ **Deprivation category**: maps dissemination areas (geographical units linked to postal code) to material and social quintiles [29] *(for future use in late 2019)* ▪ **Disease case and disease case indicator tables**: CPCSSN has developed and validated case definitions based on combinations of diagnoses, billing codes, medications and lab results; patients are indexed according to these definitions. Definitions for conditions currently include: diabetes, hypertension, depression, osteoarthritis, chronic obstructive pulmonary disease, dementia, parkinsonism, epilepsy, plus chronic kidney disease, herpes zoster, and pediatric asthma [5, 6, 14] CPCSSN redacts several types of text that could potentially identify patients, providers, or clinics. ▪ Non-medical words that may identify specific CPCSSN patients and providers are compared against exclusion lists, containing common words and/or allergy, diagnosis, and/or medication-specific terms to determine which are to be redacted. The exclusion list(s) to apply depend on the type of data being deidentified. Only those names not in the exclusion list(s) will be redacted in free text. ▪ Other identifiers, such as social insurance numbers, province health numbers, driver's license numbers, credit card numbers, Worker's Compensation Board numbers, email addresses, and phone numbers (international, North American, and local) are suppressed
Data processing validation routines	During **extractions**, the CPCSSN data are verified by: ▪ Comparing the number of files and data volume between extraction cycles occurring every six months ▪ Patient and provider counts ▪ Table row counts ▪ Counts for CPCSSN-specific labs and exams During the **data transformation**, a time-stamped log-file provides an itemized description of all processing steps completed, including: ▪ Initial and final row counts for all tables ▪ Type(s) of transformation(s) and number of data elements transformed ▪ Warnings and error messages (e.g. tables missing, processes that cannot be completed) During the **coding and cleaning processes**, the following validation tests are performed: ▪ Time-stamped log-file which includes an itemized list of completed processes and warnings where data should be coded or is outside allowable ranges ▪ Coder process verification tests used to compare the effect of software changes on specific coding or cleaning processes (e.g. ATC coding, exam data cleaning, deidentification) pre- and post-change ▪ End-to-end test for case definition integration that compares changes in the validity metrics (i.e. sensitivity, specificity) for CPCSSN case definitions after any changes to any part of the coding and cleaning software

Audit trail	Separate time- and date-stamped log files are retained for each of the extraction, transformation, and cleaning/coding stages. ▪ **Extraction**: one file per clinic; contains lists of tables and columns extracted and any warnings or errors that occurred (e.g. connection lost, table missing). An introspection table is created detailing the properties of the source database (e.g. table/column names, data types). ▪ **Transformation**: one file per clinic; contains software version information, file locations, initial and final row counts for all tables, type(s) of transformation(s) and number of data elements transformed, warnings and error messages (e.g. tables missing, processes that cannot be completed) ▪ **Coding/Cleaning**: one per PBRN; includes an itemized list of processes completed, warnings where data cannot be coded or is outside of allowed ranges