BMJ Health & Care Informatics

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

Stephanie Garies,[1,2] Erik Youngson,[3] Boglarka Soos,[1,2] Brian Forst,[4] Kimberley Duerksen,[4] Donna Manca,[4] Kerry McBrien,[1,2] Neil Drummond,[1,4,5] Hude Quan,[2] Tyler Williamson[2]

To cite: Garies S, Youngson E, Soos B, *et al*. Primary care EMR and administrative data linkage in Alberta, Canada: describing the suitability for hypertension surveillance. *BMJ Health Care Inform* 2020;**27**:e100161. doi:10.1136/bmjhci-2020-100161

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2020-100161).

Received 15 April 2020
Revised 09 June 2020
Accepted 29 June 2020

[1]Family Medicine, University of Calgary, Calgary, Alberta, Canada
[2]Community Health Sciences, University of Calgary, Calgary, Alberta, Canada
[3]Alberta Strategy for Patient Oriented-Research (SPOR) SUPPORT Unit Data Platform, University of Alberta, Edmonton, Alberta, Canada
[4]Family Medicine, University of Alberta, Edmonton, Alberta, Canada
[5]School of Public Health, University of Alberta, Edmonton, Alberta, Canada

Correspondence to
Stephanie Garies;
sgaries@ucalgary.ca

ABSTRACT

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

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

Results Data were successfully linked for 6307 hypertensive patients (96.2% of eligible patients) from 49 contributing providers. Non-linkages from invalid PHN (n=246) occurred more for deceased patients and those with fewer primary care encounters, with differences due to type of EMR and patient EMR status. The linked cohort had more patients who were female, >60 years and residing in rural areas compared to the provincial healthcare registry. Family physicians were more often female and medically trained in Canada compared to all physicians in Alberta. Most patients (>97%) had ≥1 record in the registry, pharmacy, emergency/ambulatory care and claims databases; 44.3% had ≥1 record in the hospital discharge database.

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

Summary box

What is already known?

- Data linkage can enhance existing chronic disease surveillance systems by providing more comprehensive, detailed information about patients throughout the healthcare system.
- Linking data from electronic medical records (EMR) and administrative sources is not routinely conducted in many regions in Canada.

What does this paper add?

- Describes the process for linking primary care EMR data with administrative data in Alberta, Canada.
- Evaluates the linkage results and assesses the linked database for use in hypertension surveillance.

INTRODUCTION

Chronic disease surveillance is an important public health function, which includes monitoring health events over time and examining processes of disease such as aetiology, treatment patterns, long-term management and health outcomes.[1–3] In order to be effective, a robust surveillance system should include information that is current and timely, comprehensive, accurate, accessible at various levels, longitudinal, stable, flexible, cost-effective and population-based or representative of the population.[1 4 5]

In Canada, administrative data are often used for chronic disease surveillance, as it is routinely collected, longitudinal information on healthcare encounters for nearly the entire population.[3 6] The proliferation of electronic medical record (EMR) systems in healthcare settings has now established EMR data as another option for disease surveillance.[7 8] Both sources are not without limitations: administrative data lack important clinical indicators and risk factors, while EMR data are extracted for a small subset of the population and are more difficult to analyse.

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

health information laws, data holdings and data accessibility for secondary uses.

We used primary care EMR data from one Canadian province (Alberta) to conduct linkages with five administrative databases for patients with hypertension. Hypertension was chosen as an example chronic condition, as it has been identified as an important surveillance priority and is largely managed in primary care.[9 10] Our objective was to describe the current process for EMR-administrative data linkage in Alberta and to examine the advantages, potential biases and limitations of using linked data for hypertension surveillance.

METHODS

Data sources

Primary care EMR data

The Canadian Primary Care Sentinel Surveillance Network (CPCSSN) is a national repository of de-identified primary care EMR data for nearly two million patients which is available for approved secondary purposes such as research and surveillance.[11 12] CPCSSN is a collaboration of practice-based research networks in seven Canadian provinces. Each regional network extracts, processes and standardises EMR data from participating family physicians and nurse practitioners (all termed 'sentinels') which is then submitted to the national repository. The CPCSSN data include patient demographics, diagnoses, prescribed medications, physical measurements (eg, blood pressure, height, weight), physician billing claims, behavioural risk factors, laboratory results, referrals and medical procedures. The national CPCSSN database includes approximately 4.8% of all Canadians as of December 2019[13] and has been found to be slightly over-represented by females and older adults, though this is consistent with primary care populations.[14] Family physicians who contribute data to CPCSSN constitute an estimated 3.3% of all family physicians in Canada.[15]

Alberta, 1 of 13 Canadian provinces and territories, is located in the western part of the country with a population of just under 4.4 million people in 2019.[13] In Alberta, the Northern and Southern Alberta Primary Care Research Networks (NAPCReN and SAPCReN, respectively) extract de-identified EMR data twice annually from five commonly used EMR systems. Longitudinal patient data are available dating back to the start of EMR use, which varies by clinic and by patient, until the most recent data extraction (up to 30 June 2018 for this study). Included in this sample were data for 204 825 adult patients and 310 primary care providers, representing 4.8% and 5.6% of the total Alberta citizens and family physicians, respectively.[15 16] Patients who have explicitly opted out of the CPCSSN database are excluded. There are more females, older adults and rural patients in the CPCSSN database for Alberta compared to the provincial healthcare registry which is typically used as a population denominator (online supplementary appendix 1).

Additional details about the CPCSSN data extraction and processing in Alberta have been published previously.[17]

Administrative data

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

1. Discharge Abstract Database (DAD): in-patient hospital discharges from all acute care institutions in Alberta. Each record contains the most responsible diagnosis and up to 24 secondary diagnosis fields coded using the International Classification of Diseases version 10, Canadian Enhancement (ICD-10-CA).
2. National Ambulatory Care Reporting System (NACRS): ambulatory care visits and procedures (eg, day surgery, emergency room visits, community rehabilitation services). NACRS contains a primary diagnosis with up to nine secondary codes for each record using ICD-10-CA.
3. Practitioner claims: submitted billing claims from fee-for-service physicians and other allied health practitioners, in addition to shadow-billed claims from alternative payment models, with information about the type of provider and service provided.
4. Pharmaceutical Information Network (PIN): dispensed prescription medications from approximately 99% of community pharmacists in Alberta. Details include drug name/identification number, dose, quantity dispensed and supply amount. PIN does not capture medications dispensed in hospitals or emergency departments.
5. Alberta Health Care Insurance Plan (AHCIP) registry: population-level demographic registry of all Albertans registered for publicly funded healthcare in the province; includes nearly every resident except members of Canadian Armed Forces and inmates of federal penitentiaries (who are instead covered by the federal government).

Data linkage

Figure 1 outlines the current process for linking de-identified EMR data from CPCSSN in Alberta with identifiable administrative data from AHS. CPCSSN sentinels were required to formally agree to the data linkage through a research agreement specific to this project. Sentinels who agreed to the data linkage were asked to generate an EMR Mapping File from their clinic EMR that included patients' personal healthcare number (PHN), EMR ID, sex and date of birth. PHN is a unique identifier assigned to residents of Alberta who have registered with the AHCIP. Additionally, CPCSSN generated a 'CPCSSN' Mapping File that contained two variables for each hypertensive patient in the CPCSSN database belonging to a participating sentinel: EMR ID (collected at time of CPCSSN's routine data extraction) and CPCSSN ID

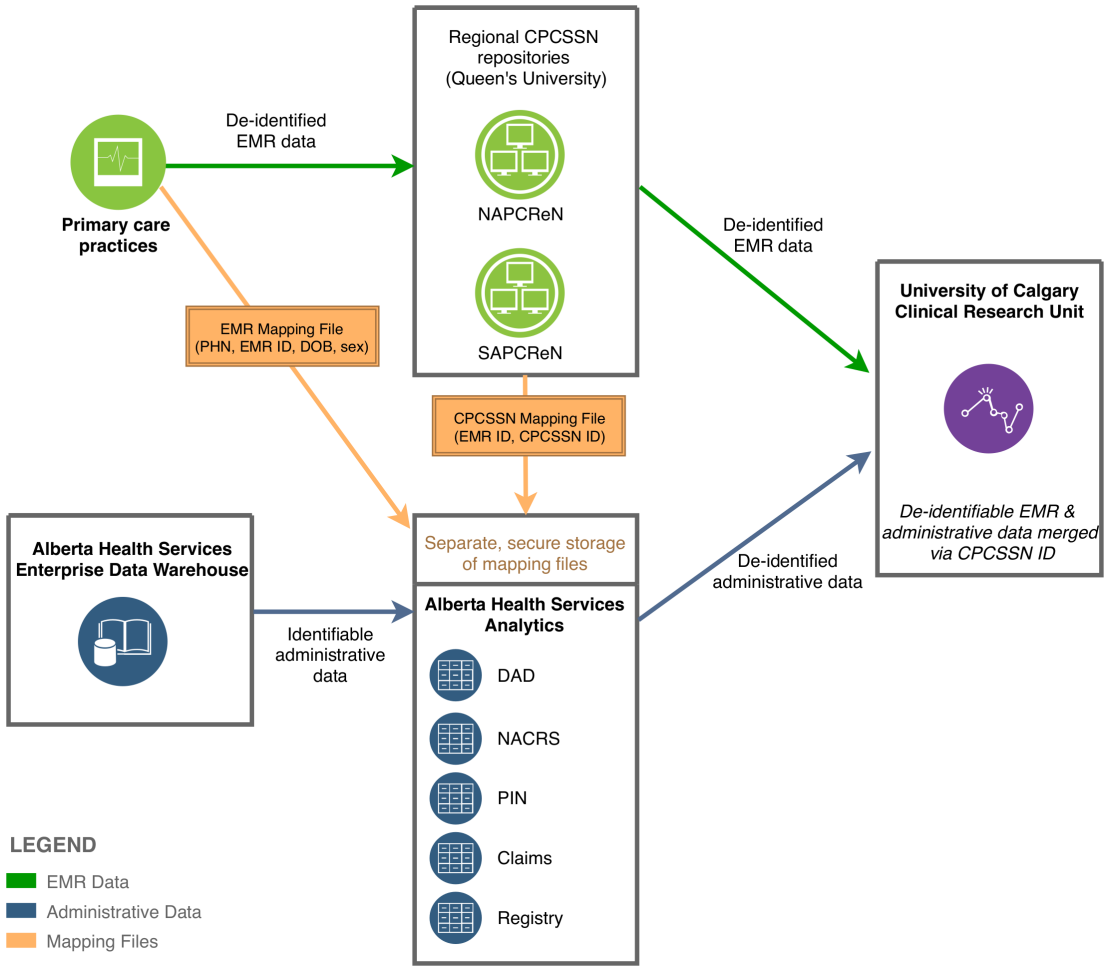

Figure 1 Data flow and linkage process for primary care EMR and administrative data in Alberta. CPCSSN, Canadian Primary Care Sentinel Surveillance Network; DAD, Discharge Abstract Database; EMR, electronic medical record; PHN, personal healthcare number; PIN, Pharmaceutical Information Network; NACRS, National Ambulatory Care Reporting System.

(a unique, randomly assigned study ID). The EMR and CPCSSN Mapping Files were securely transferred to AHS Analytics.

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

Defining hypertension

Patients with hypertension were identified using validated case definitions for both data sources (table 1). The administrative definition is based on two physician claims using relevant ICD-9 codes within 2 years or one hospitalisation using ICD-10 codes within the DAD.[18] The EMR-based definition was developed by CPCSSN and uses a combination of text words, ICD-9 diagnostic codes and antihypertensive medications located throughout the EMR; when validated, this definition performed slightly better than the administrative definition.[19]

Analysis

The overall linkage rate was calculated as the proportion of eligible hypertensive patients in the CPCSSN cohort that matched exactly by PHN (and one of sex or date of birth) to the AHCIP registry. Patient and clinic characteristics for those who were in the linked cohort and those who could not be linked were compared using a chi-squared test for categorical data, t-test for continuous variables and Wilcoxon test for non-parametric data.

Selected demographic characteristics of adult patients in the linked hypertensive cohort were reported and compared with characteristics for all hypertensive patients in the full CPCSSN data for Alberta. Characteristics of family physicians contributing linked data were compared with all physicians in the CPCSSN database in Alberta and to all family medicine physicians in Alberta according to the Canadian Institute for Health Information *Supply, Distribution and Migration of Physicians in Canada* report.[15]

The number of patients meeting any of the criteria from the administrative and EMR definitions for hypertension

Table 1 Criteria for hypertension definitions in administrative and EMR data

Definition	Data source	Criteria	Codes	Validity
Administrative[18]	Physician claims	At least two billing claims within 2 years	ICD9 codes: 401–405	Sensitivity: 75% Specificity: 94% PPV: 81% NPV: 92%
	Discharge Abstract Database	One in-patient diagnosis code at any time	ICD10 codes: I10-I15	
EMR/CPCSSN[19]	Physician claims table	At least two billing claims within 2 years	ICD9 codes: 401–405	Sensitivity: 84.9% Specificity: 93.5 PPV: 92.9% NPV: 86.0%
	Problem list/ profile table	At least one diagnosis code at any time	ICD9 codes: 401–405	
	Prescribed medication table	Any occurrence of a specified hypertension medication, *except* if one of the following co-morbid conditions exists: diabetes mellitus (250), tremor (333.1), migraine (346), myocardial infarction (410, 412), angina or cardiac dysrhythmias (413, 427), heart failure (428), oesophageal varices (456.0, 456.1), calculus of kidney and ureter (592), portal hypertension (572.3)	ATC codes: C02, C03AA03, C03BA04, C03BA08, C03BA11, C03DB01, C03DB02, C03EA01, C07AA06, C07AB03, C07AB04, C07AG02, C07CB03, C08CA01, C08CA02, C08DA01, C09AA01, C09AA02, C09AA03, C09AA07, C09AA08, C09AA09, C09AA10, C09BA02, C09BA03, C09CA02, C09CA03, C09CA04, C09CA07, C09DA01, C09DA02, C09DA04, C09XA02	

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

was explored by sex and age group. All data analyses were performed in RStudio V.1.1.456.

RESULTS

Data linkage

Figure 2 presents the flow of patients throughout the linkage process. Within the CPCSSN database, 50 342 adult patients were identified as having hypertension (24.5%). Fifty-five sentinels out of a possible 243 CPCSSN sentinels across Alberta had explicitly agreed to the data linkage, however, we were unable to link data for six sentinels (two were excluded due to data extraction issues with the EMR vendor; four providers were unable to have their EMR Mapping File submitted from the clinic). From the remaining 49 sentinels, a total of 6553 patients were eligible for linkage with administrative data.

Overall, we were able to link 6307 of the 6553 (96.2%) eligible patients from the CPCSSN EMR cohort to the administrative data. The proportion of patients in the CPCSSN data who had at least one record in each of the administrative databases was very high (100% for AHCIP Registry, PIN and Claims; 97.2% for NACRS) or expected (44.3% for DAD in-patient hospitalisations). Of those whose data were unsuccessfully linked, the majority (n=200; 81.3%) was due to having no available PHN; the remaining 18.7% (n=46) of unlinked patients had mismatched sex and/or date of birth between the EMR Mapping File and AHCIP Registry.

Table 2 describes selected characteristics of patients who had their EMR data linked to administrative data and those whose data could not be linked. Patients with unlinked CPCSSN records were more likely to be deceased according to the EMR (24.4% compared to 0.5% in the linked cohort; p<0.001) and were also associated with a lower mean number of primary care encounters (39.2 compared to 45.8 in the linked cohort; p=0.013). Patients with linked records also differed by EMR status (p<0.001) and the type of EMR system used at the practice (p<0.001) compared to those with unlinked records.

Patient cohort comparisons

Table 3 describes demographic characteristics of adult patients with hypertension in the linked cohort (n=6307) compared to all adults with hypertension in the full CPCSSN dataset from Alberta (n=50 342). The linked cohort had slightly more females (53.8% vs 51.4%), similar mean ages (64.7 vs 65.4 years) and fewer urban residents (69.5% vs 76.1%) than all patients with hypertension in the full CPCSSN database for Alberta. The full CPCSSN database compared to the healthcare registry (population denominator) was observed to be more over-represented by females (56.3%), older patients within several age bands and rural patients (19.2% vs 9.0%) (online supplementary appendix 1).

Provider comparisons

Table 4 describes the characteristics of family physicians who completed the data linkage (n=48) compared to two groups: (1) all family medicine physicians in the province; and (2) family physician sentinels contributing data to CPCSSN dataset within Alberta. The linked cohort

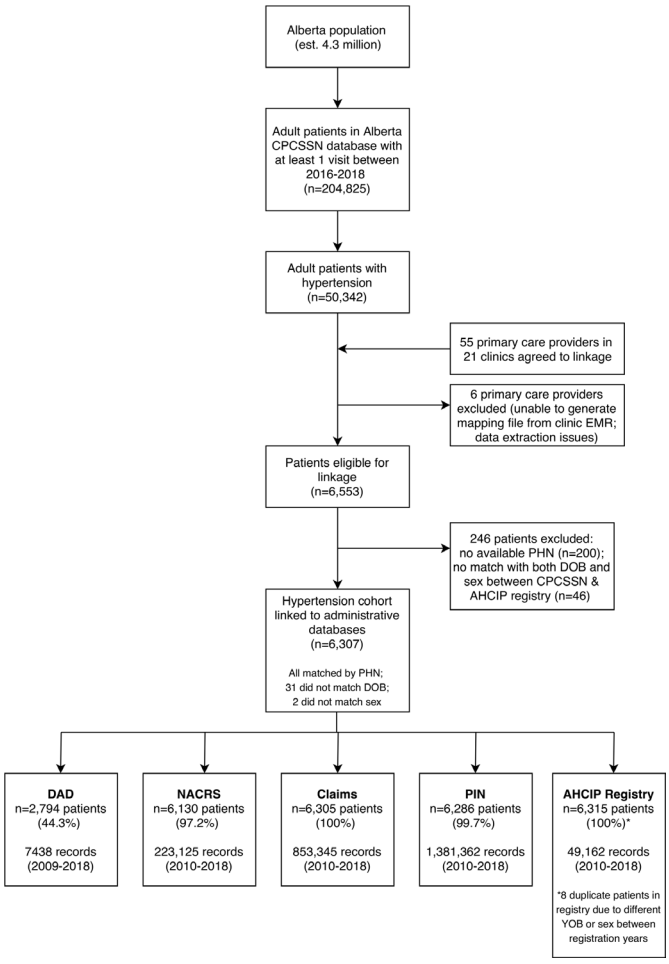

Figure 2 Flow diagram of patient selection into the linked hypertension cohort. AHCIP, Alberta Health Care Insurance Plan; CPCSSN, Canadian Primary Care Sentinel Surveillance Network; DAD, Discharge Abstract Database; DOB, date of birth; EMR, electronic medical record; NACRS, National Ambulatory Care Reporting System; PHN, personal healthcare number; PIN, Pharmaceutical Information Network; YOB, year of birth.

included more physicians who were female (56.2% vs 42.7%), fewer who were internationally trained (2.1% vs 43.2%) and slightly less from rural practices (10.4% vs 12.2%) when compared with all family physicians in Alberta.[15] Although the mean age was similar, the linked cohort had higher proportions of physicians between the ages of 30–39 years and 60–69 years. Compared to all participating CPCSSN sentinels in Alberta, family physicians contributing to the linkage were observed to be slightly younger (mean age 48.7 years in linked cohort vs 47.1) with lower proportions of internationally trained physicians and those practising in rural locations.

Hypertension definitions

Overall, there were 1276 patients (20.2%) that met all five criteria from both the administrative and EMR definitions; nearly 9% of patients (n=564) met at least one component of the EMR hypertension definition, but did not meet any administrative criteria (data not shown).

Table 2 Comparison of characteristics for hypertensive patients with and without linked data

	Linked (n=6307)	Not linked (n=246)	P value
Female, n (%)	3395 (53.8)	127 (51.6)	0.539
Age, mean (SD)	64.7 (14.1)	65.2 (17.2)	0.523
Deceased year recorded in the EMR, n (%)	30 (0.5)	60 (24.4)	<0.001
Patient EMR status			<0.001
Active, n (%)	5326 (84.4)	126 (51.2)	
Deceased, n (%)	104 (1.6)	60 (24.4)	
Inactive, n (%)	99 (1.6)	8 (3.3)	
Unknown, n (%)	778 (12.3)	52 (21.1)	
Urban residence, n (%)*	4312 (69.5)	156 (65.5)	0.225
Type of EMR			<0.001
Wolf, n (%)	3266 (51.8)	221 (89.8)	
Med Access, n (%)	3041 (48.2)	25 (10.2)	
Number of primary care encounters, mean (SD)	45.8 (40.7)	39.2 (47.2)	0.013

*Postal code for determining urban or rural residence was missing for 99 patients (1.6%) in the Linked cohort and eight patients (3.3%) in the Not Linked cohort.
EMR, electronic medical record.;

Table 3 Comparison of hypertensive patient characteristics in the CPCSSN database in Alberta and the linked EMR-administrative cohort

	Primary care EMR data (all patients with hypertension)	Linked admin-EMR data (hypertension cohort)
Data source	CPCSSN data in Alberta up to 30 June 2018; patients with at least one visit in last 2 years	CPCSSN data in Alberta linked to AHCIP Registry and other admin databases
Total adults, N	50342	6307
Female, n (%)	25865 (51.4)	3395 (53.8)
Male, n (%)	24475 (48.6)	2912 (46.2)
Age, mean (SD)	65.4 (14.1)	64.7 (14.1)
Age groups, n (%)		
20–39 years	2247 (4.5)	330 (5.2)
40–59 years	14020 (27.8)	1790 (28.4)
60–69 years	14030 (27.9)	1833 (29.1)
70–79 years	11662 (23.2)	1397 (22.1)
80 years and older	8383 (16.7)	957 (15.2)
Urban, n (%)	37507 (76.1)	4312 (69.5)
Rural, n (%)	11783 (23.9)	1896 (30.5)
Missing/unknown residence or postal code, n (%)	1053 (2.1)	99 (1.6)

AHCIP, Alberta Health Care Insurance Plan; CPCSSN, Canadian Primary Care Sentinel Surveillance Network; EMR, electronic medical record.;

Table 4 Comparison of family physician characteristics in Alberta and in the CPCSSN data

	All family medicine physicians in Alberta[15]	Family physicians contributing to CPCSSN AB (data up to June 2018)	Family physicians who agreed to EMR-admin data linkage (hypertension cohort)
Total, N	5489	310	48*
Female, n (%)	2343 (42.7)	171 (55.2)	27 (56.2)
Male, n (%)	3146 (57.3)	139 (44.8)	21 (43.8)
Mean age, years (SD)	48.8	47.1 (10.3)	48.7 (12.0)
<30, n (%)	217 (4.0)	1 (0.3)	0 (0)
30–39, n (%)	1354 (24.7)	67 (21.6)	13 (31.7)
40–49, n (%)	1441 (26.3)	79 (25.5)	9 (22.0)
50–59, n (%)	1191 (21.7)	72 (23.2)	7 (17.1)
60–69, n (%)	542 (9.9)	35 (11.3)	11 (26.8)
70+, n (%)	731 (13.3)	3 (1.0)	1 (2.4)
Missing age, n (%)	13 (0.2)	53 (17.1)	7 (14.6)
Rural practice, n (%)	670 (12.2)	43 (13.9)	5 (10.4)
International medical training, n (%)	2373 (43.2)	69 (22.3)	1 (2.1)
Missing location of medical training, n (%)	n/a	19 (6.1)	1 (2.1)

*Excludes one nurse practitioner from the 49 providers who contributed to the linkage.
AB, Alberta; CPCSSN, Canadian Primary Care Sentinel Surveillance Network; EMR, electronic medical record.;

From the two administrative criteria, most patients met the billing claims criterion; from the three EMR criteria, the EMR profile was the most frequently met criterion (n=4801), though the billing claims (n=4693) and prescribed medication (n=4449) also had similar numbers of patients (figure 3). Major discrepancies between definitions by sex and age group were not evident.

DISCUSSION

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

The advantages and limitations of a linked EMR-administrative cohort with respect to important features and information needs of a surveillance system (box 1) will be discussed.

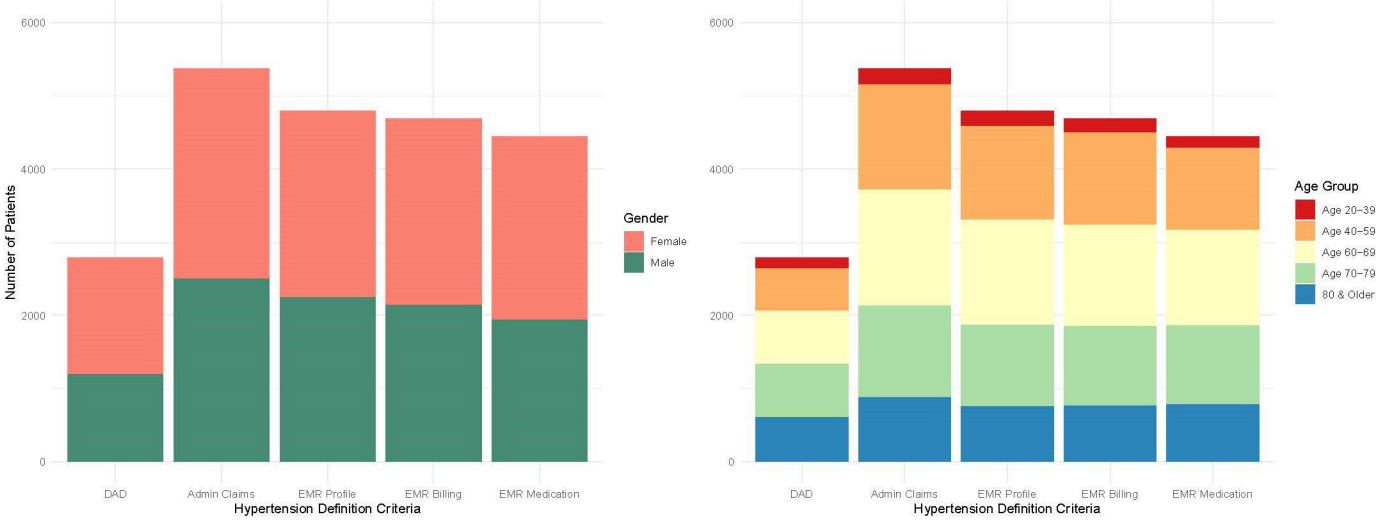

Figure 3 Patients in the linked data cohort meeting hypertension case criteria for administrative and EMR definitions by sex and age group (n=6307). DAD, Discharge Abstract Database; EMR, electronic medical record.

Box 1 Criteria for a robust chronic disease surveillance system

System features[1 4 5]
- Current, timely.
- Comprehensive.
- Accurate.
- Accessible at various levels.
- Longitudinal.
- Stable.
- Flexible.
- Cost-effective.
- Population-based or representative of the population.

Information needs for hypertension surveillance
- Biological risk factors (eg, blood pressure, obesity/body mass index, pre-existing conditions, cholesterol & other lab markers).
- Lifestyle risk factors (eg, smoking status, physical inactivity, alcohol use, diet).
- Disease progression (eg, lab results, severity of disease).
- Disease management (eg, medications prescribed and dispensed; medical procedures).
- Health outcomes (eg, subsequent diagnoses, acute events requiring hospitalisation, mortality).
- Environmental and policy factors.
- Sociodemographics (eg, ethnicity, occupation, poverty).
- Patient-reported outcome measures and experiences (eg, symptom burden, functional status, quality of life).
- Healthcare costs.

Advantages

Comprehensiveness

The most significant advantage of this type of data linkage is the creation of more comprehensive information about people with hypertension, more so than either data source on its own. For instance, smoking status, weight trajectories over time, medication prescriptions and longitudinal blood pressure measurements recorded in primary care practices are now available alongside pharmacy-dispensed medication records and acute outcomes requiring hospitalisation, which does not exist currently in Alberta. This combination of community-based, detailed clinical information and system-wide healthcare encounters can help to enhance our understanding of disease processes, management and outcomes.

Completeness

Linking EMR and administrative data can also help to improve data completeness and evaluate the accuracy of data elements. The Electronic Medical Record Administrative data Linked Database in Ontario, Canada has demonstrated this in their assessment of prescribed medication validity in EMRs compared to a provincial drug database, as well as using linked data to confirm a variety of administrative-based disease identification algorithms.[20–23]

Timeliness

Both administrative and EMR data are available in a timeframe suitable for chronic disease surveillance, as real-time data are not necessarily warranted for conditions that have a long latency period. However, the linkage process described here took just over 1 year to complete, after research ethics approvals, obtaining sentinel agreements to link EMR data, generating the mapping files at each of the participating clinics, and completing the EMR and administrative linkage. Even so, updating linkages once per year, for example, would still be reasonable for hypertension or other chronic disease surveillance.

Limitation and potential biases

Coverage of capture

The primary care setting represents citizens who have accessed the first level of the healthcare system and the CPCSSN database is a further subset of these patients whose providers consented to participate. The CPCSSN data from Alberta were observed to include higher proportions of physicians who were female, completed medical training in Canada and practice in rural locations. Patients in the Alberta CPCSSN data were more likely to be female, older and located in rural areas compared to all AHCIP registrants, although this reflects a typical primary care population in terms of older adults and women.[14]

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

Completeness and accuracy

Although this linkage brought together previously discrete data about patients with hypertension, risk factors and other information relevant to hypertension surveillance are still missing or incomplete. For example, smoking status and alcohol use are often entered inconsistently as free text in EMRs and may not be recorded for all patients in an extractable format, which limits their usability.[24 25] Environmental and socioeconomic factors also contribute to the aetiology and management of hypertension; however, neither primary care EMR nor administrative health data capture this information sufficiently. This highlights the need for additional data linkage to occur with multiple health and non-health sources (eg, ministerial data from occupation, education, social services; surveys; geographic information systems (GIS) mapping), while still respecting individual privacy and maintaining minimal risk of re-identification.

There were a few issues with the accuracy of the linkage process. Duplicate patients were found in the AHCIP Registry due to different year of birth or sex recorded between subsequent registration years; however, this represented a very small proportion of patients (0.1%).

Another challenge related to the process of generating the mapping files. Because CPCSSN extracts data from five different EMR systems in the province, there are variations in the EMR data extraction methods (which also generates the CPCSSN mapping files) and creation of the EMR mapping files from each clinic, as well as differences in how patients are assigned (or not assigned) to providers in the EMR. This was evident in the differing linkage rates by EMR type, by deceased status (which is not necessarily known or recorded in primary care settings) and in the EMR status of patients in the clinic (eg, whether only active or all patients were included in the mapping files). The process for extracting the mapping files in the clinic EMR and the criteria for patient selection requires further standardisation.

Misclassification

Misclassifying those who truly have or do not have hypertension by the case finding algorithms can greatly bias epidemiological or surveillance outcomes. Previous validation studies have shown that the EMR-based definition for hypertension outperformed the administrative definition in both sensitivity (84.5% vs 75%) and PPV (92.9% vs 81%).[18 19] In this cohort, 564 (8.9%) of the hypertensive patients in the EMR were not identified as hypertensive in the administrative data; this may be due to a longer timeframe of available data in the EMR or less frequent hypertension billing claims submitted for patients with long-term control.

Poor data quality can also contribute to misclassification through data entry errors in the clinic EMR from the use of non-specific or no diagnostic codes, information recorded in inaccessible formats (eg, free text notes; scanned documents) or errors produced during CPCSSN processing.[26] Since the administrative definition for hypertension only uses the ICD-9 billing codes and a smaller number of ICD-10-CA codes within in-patient data, this may further limit the ability of the administrative data to identify all true cases of hypertension and thus, linkage with detailed EMR data can augment case definition accuracy.

CONCLUSION

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

Contributors All authors contributed to the study conception and design. SG, EY, BF, BS and KD supported and/or conducted the data linkage. SG conducted the analysis, wrote the first draft of the manuscript, and managed revisions. All authors read and approved the final manuscript.

Funding In-kind support was provided by the Alberta Strategy for Patient Oriented Research (SPOR) SUPPORT Unit Data Platform to facilitate data linkage with the Alberta Health Services Enterprise Data Warehouse. SG is funded through an Alberta Innovates Health Solutions Graduate Studentship (2016–2020). The CPCSSN project in Alberta, hosted by NAPCReN and SAPCReN, receives funding from the Canadian Institutes of Health Research (CIHR) and Alberta Innovates through the Alberta Strategies for Patient Oriented Research (SPOR) Primary and Integrated Health Care Innovation Network, as well as the Public Health Agency of Canada.

Disclaimer The funders had no role in the study design, data collection, analysis or interpretation of the data, or in the writing of the manuscript.

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval As part of the CPCSSN project, both NAPCReN and SAPCReN have research ethics approval to extract de-identified EMR data through their respective universities. This linkage study was granted separate approval by the Conjoint Health Research Ethics Board at the University of Calgary (REB17-1825) and the Health Research Ethics Board at the University of Alberta (Pro00079372).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. The Alberta-specific CPCSSN data that were used for this work are available as two separate data sets through the regional networks (NAPCReN, SAPCReN). Data access procedures and requirements vary by network; contact the corresponding author for more information or visit: http://napcren.ca/, http://sapcren.ca. The national CPCSSN data are available to approved researchers for a fee; for more information or to submit a Letter of Intent for data access, visit: http://cpcssn.ca/research-resources/.

REFERENCES

1 Institute of Medicine. *A nationwide framework for surveillance of cardiovascular and chronic lung diseases*. Washington DC: The National Academies of Press, 2011.
2 Centre for Surveillance Coordination. *Chronic disease surveillance in Canada: a background paper*. Ottawa, Canada, 2003.
3 Lix LM, Yogendran MS, Shaw SY, *et al*. Population-based data sources for chronic disease surveillance. *Chronic Dis Can* 2008;29:31–8.
4 German RR, Lee LM, Horan JM, *et al*. Updated guidelines for evaluating public health surveillance systems: recommendations from the guidelines Working group. *MMWR Recomm Rep* 2001;50:1–35.
5 Goff DC, Brass L, Braun LT, *et al*. Essential features of a surveillance system to support the prevention and management of heart disease and stroke. *Circulation* 2007;115:127–55.
6 Public Health Agency of Canada, Canadian Chronic Disease Surveillance System. Public health infobase: Canadian Chronic Disease Surveillance System (CCDSS) [Internet], 2017. Available: https://infobase.phac-aspc.gc.ca/ccdss-scsmc/data-tool/
7 Birkhead GS, Klompas M, Shah NR. Uses of electronic health records for public health surveillance to advance public health. *Annu Rev Public Health* 2015;36:345–59.
8 Birtwhistle R, Williamson T. Primary care electronic medical records: a new data source for research in Canada. *CMAJ* 2015;187:239–40.
9 Campbell NRC, McAlister FA, Quan H, *et al*. Monitoring and evaluating efforts to control hypertension in Canada: why, how, and what it tells us needs to be done about current care gaps. *Can J Cardiol* 2013;29:564–70.
10 Godwin M, Williamson T, Khan S, *et al*. Prevalence and management of hypertension in primary care practices with electronic medical records: a report from the Canadian primary care sentinel surveillance network. *CMAJ Open* 2015;3:E76–82.
11 CPCSSN. Canadian Primary Care Sentinel Surveillance Network (CPCSSN) [Internet], 2016. Available: http://www.cpcssn.ca/

12 Garies S, Birtwhistle R, Drummond N, *et al*. Data resource profile: national electronic medical record data from the Canadian primary care sentinel surveillance network (CPCSSN). *Int J Epidemiol* 2017;46:1091–2.
13 Statistics Canada. Population estimates, quarterly [Internet]. Table 17-10-0009-01, 2020. Available: https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1710000901
14 Queenan JA, Williamson T, Khan S, *et al*. Representativeness of patients and providers in the Canadian primary care sentinel surveillance network: a cross-sectional study. *CMAJ Open* 2016;4:E28–32.
15 Canadian Institute for Health Information. Supply, distribution and migration of physicians in Canada, 2018 - data tables [Internet], 2018. Available: https://secure.cihi.ca/estore/productSeries.htm?pc=PCC34
16 Government of Alberta. Quarterly population report; Second quarter 2019 [Internet], 2019. Available: https://open.alberta.ca/dataset/aa3bce64-c5e6-4451-a4ac-cb2c58cb9d6b/resource/ae2c77eb-3ce0-4f70-90a3-6a2329f49355/download/2019-q2-population-report.pdf
17 Garies S, Cummings M, Forst B, *et al*. Achieving quality primary care data: a description of the Canadian primary care sentinel surveillance network data capture, extraction, and processing in Alberta. *Int J Popul Data Sci* 2019;4.
18 Quan H, Khan N, Hemmelgarn BR, *et al*. Validation of a case definition to define hypertension using administrative data. *Hypertension* 2009;54:1423–8.
19 Williamson T, Green ME, Birtwhistle R, *et al*. Validating the 8 CPCSSN case definitions for chronic disease surveillance in a primary care database of electronic health records. *Ann Fam Med* 2014;12:367–72.
20 Schwartz KL, Wilton AS, Langford BJ, *et al*. Comparing prescribing and dispensing databases to study antibiotic use: a validation study of the electronic medical record administrative data linked database (EMRALD). *J Antimicrob Chemother* 2019;74:2091–7.
21 Schultz SE, Rothwell DM, Chen Z, *et al*. Identifying cases of congestive heart failure from administrative data: a validation study using primary care patient records. *Chronic Dis Inj Can* 2013;33:160–6.
22 Tu K, Wang M, Jaakkimainen RL, *et al*. Assessing the validity of using administrative data to identify patients with epilepsy. *Epilepsia* 2014;55:335–43.
23 Tu K, Wang M, Young J, *et al*. Validity of administrative data for identifying patients who have had a stroke or transient ischemic attack using EMRALD as a reference standard. *Can J Cardiol* 2013;29:1388–94.
24 Greiver M, Aliarzadeh B, Meaney C, *et al*. Are we asking patients if they smoke?: missing information on tobacco use in Canadian electronic medical records. *Am J Prev Med* 2015;49:264–8.
25 Torti J, Duerksen K, Forst B, *et al*. Documenting alcohol use in primary care in Alberta. *Can Fam Physician* 2013;59:1128.
26 Coleman N, Halas G, Peeler W, *et al*. From patient care to research: a validation study examining the factors contributing to data quality in a primary care electronic medical record database. *BMC Fam Pract* 2015;16:11.