Assessing the quality of primary care electronic health record data in Australia and Canada: a case-study in Osteoarthritis

Sharmala Thuraisingam, Dewdunee Himasara Marasinghe, Kendra Barrick, Fariba Aghajafari, Jo-Anne Manski-Nankervis, Michelle Maree Dowsey, Hude Quan, Tyler Williamson, Stephanie Garies

Submitted to: Journal of Medical Internet Research
on: December 04, 2024

Table of Contents

Assessing the quality of primary care electronic health record data in Australia and Canada: a case-study in Osteoarthritis

Sharmala Thuraisingam[1]; Dewdunee Himasara Marasinghe[2]; Kendra Barrick[3]; Fariba Aghajafari[2]; Jo-Anne Manski-Nankervis[4]; Michelle Maree Dowsey[1]; Hude Quan[5]; Tyler Williamson[5]; Stephanie Garies[2]

[1]Department of Surgery University of Melbourne Melbourne AU
[2]Department of Family Medicine Cumming School of Medicine University of Calgary Calgary CA
[3]Ridgeview Medical Centre Canmore CA
[4]Family Medicine and Primary Care Lee Kong Chian School of Medicine Nanyang Technological University Singapore SG
[5]Department of Community Health Sciences Cumming School of Medicine University of Calgary Calgary CA

Corresponding Author:
Sharmala Thuraisingam
Department of Surgery
University of Melbourne
29 Regent Street, Fitzroy, Victoria 3065, Australia
Melbourne
AU

Abstract

Background: General practice electronic health records contain a wealth of patient information. However, these data were collected for clinical purposes. Hence, questions remain around the suitability of using these data for other purposes including epidemiological research, developing and validating clinical prediction models, conducting audits and informing policy. This study assessed the quality of data in Australian and Canadian general practice electronic health records in the context of osteoarthritis for the purpose of externally validating a clinical prediction model for total knee replacement surgery.

Objective: To assess the quality of data in Australian and Canadian primary care electronic health records for the purpose of externally validating a clinical prediction model for use in patients with osteoarthritis.

Methods: A data quality assessment was conducted on 201,462 patient general practice electronic health records from Australia provided by NPS MedicineWise, and 92,425 from Canada provided by the Canadian Primary Care Sentinel Surveillance Network. Completeness, plausibility and external validity of data elements relevant to osteoarthritis were assessed.

Results: There were minimal incomplete and implausible data fields for age and gender (<1%), geographical location (<5%) and commonly co-occurring comorbidities (<10%) in both data sets. However, weight, height, body mass index and Canadian Index of Multiple Deprivation contained over 50% missing data. The recording of osteoarthritis by age and gender in both data sets were similar to national estimates, except for patients aged 80+ (Australia: 16.6% [95% CI 16.0%-17.3%] vs 13.1% [95% CI 11.2%-15.4%]; Canada: 36.7% [95% CI 36.1%-37.2%] vs 50.8% [95% CI 50.7%- 50.9%]. Total knee replacement rates were substantially lower in both electronic health record data sets compared with national estimates (Australia: 72 vs 218 per 100,000; Canada: 0.84 vs 200 per 100,000).

Conclusions: Age, gender, geographical location, commonly co-occurring comorbidities and prescribing of osteoarthritis medications in Australian and Canadian general practice electronic health records are suitable for use in clinical prediction model validation studies. However, body mass index and Canadian Index of Multiple Deprivation are unfit for use due to large proportions of missing data. Rates of total knee replacement surgery were substantially underreported and should not be used for prediction model validation. Better harmonisation of patient data across primary and tertiary care is required to improve the suitability of these data. In the meantime, data linkage with national registries and other health data sets may overcome some of the data quality challenges in general practice EHRs.

(JMIR Preprints 04/12/2024:69631)
DOI: https://doi.org/10.2196/preprints.69631

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Assessing the quality of primary care electronic health record data in Australia and Canada: a case-study in Osteoarthritis

Authors

*Corresponding author

Sharmala Thuraisingam[1*], sharmala.thuraisingam@unimelb.edu.au

Dewdunee Himasara Marasinghe[2], dhmarasi@ucalgary.ca

Kendra Barrick[3], kendra.barrick@gmail.com

Fariba Aghajafari[2], fariba.aghajafari@ucalgary.ca

Jo-Anne Manski-Nankervis[4,5], joanne.mn@ntu.edu.sg

Michelle M Dowsey[1], mmdowsey@unimelb.edu.au

Hude Quan[6], hquan@ucalgary.ca

Tyler Williamson[6,7,8], tyler.williamson@ucalgary.ca

Stephanie Garies[2], sgaries@ucalgary.ca

1 Department of Surgery, University of Melbourne, 29 Regent Street, Fitzroy, Victoria 3065, Australia

2 Department of Family Medicine, Cumming School of Medicine, University of Calgary, HSC G012-3330 Hospital Drive NW, Calgary, Alberta T2N 4N1, Canada

3 Ridgeview Medical Centre, 212-1240 Railway Avenue, Canmore, Alberta T1W 1P4, Canada

4 Department of General Practice, University of Melbourne, 780 Elizabeth Street, Parkville, Victoria 3010, Australia

5 Family Medicine and Primary Care, Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore

6 Department of Community Health Sciences, Cumming School of Medicine, University of Calgary, 3D10-3280 Hospital Drive NW, Calgary, Alberta T2N 4Z6, Canada

7 O’Brien Institute for Public Health, Cumming School of Medicine, University of Calgary, Calgary, Canada

8 Alberta Children’s Hospital Research Institute, Cumming School of Medicine, University of Calgary, Calgary, Canada

Abstract

Background

General practice electronic health records contain a wealth of patient information. However, these data were collected for clinical purposes. Hence, questions remain around the suitability of using these data for other purposes including epidemiological research, developing and validating clinical prediction models, conducting audits and informing policy. This study assessed the quality of data in Australian and Canadian general practice electronic health records in the context of osteoarthritis for the purpose of externally validating a clinical prediction model for total knee replacement surgery.

Methods

A data quality assessment was conducted on 201,462 patient general practice electronic health records from Australia provided by NPS MedicineWise, and 92,425 from Canada provided by the Canadian Primary Care Sentinel Surveillance Network. Completeness, plausibility and external validity of data elements relevant to osteoarthritis were assessed.

Results

There were minimal incomplete and implausible data fields for age and gender (<1%), geographical location (<5%) and commonly co-occurring comorbidities (<10%) in both data sets. However, weight, height, body mass index and Canadian Index of Multiple Deprivation contained over 50% missing data. The recording of osteoarthritis by age and gender in both data sets were similar to national estimates, except for patients aged 80+ (Australia: 16.6% [95% CI 16.0%-17.3%] vs 13.1% [95% CI 11.2%-15.4%]; Canada: 36.7% [95% CI 36.1%-37.2%] vs 50.8% [95% CI 50.7%- 50.9%]. Total knee replacement rates were substantially lower in both electronic health record data sets compared with national estimates (Australia: 72 vs 218 per 100,000; Canada: 0.84 vs 200 per 100,000).

Conclusions

Age, gender, geographical location, commonly co-occurring comorbidities and prescribing of osteoarthritis medications in Australian and Canadian general practice electronic health records are suitable for use in clinical prediction model validation studies. However, body mass index and Canadian Index of Multiple Deprivation are unfit for use due to large proportions of missing data. Rates of total knee replacement surgery were substantially underreported and should not be used for prediction model validation. Better harmonisation of patient data across primary and tertiary care is required to improve the suitability of these data. In the meantime, data linkage with national registries and other health data sets may overcome some of the data quality challenges in general practice EHRs.

Keywords

Electronic medical records
Electronic health records

Data quality assessment
Primary care
General practice
Family practice
Data linkage
Osteoarthritis

Introduction

Primary care electronic health records (EHRs) are a rich source of patient data (1,2). They typically contain administrative and clinical information including demographics, clinical observations, past and current diagnoses and medications, pathology and imaging results, referral letters to other healthcare providers, and billing information (2). Given patients attend general practice on average five times a year (3), data within these records are longitudinal in nature and may allow tracking of a patient’s health journey over time. The large volumes of data contained within primary care EHRs have led to these data being used for a variety of purposes, including disease surveillance, longitudinal studies, prediction modelling, auditing primary care, pharmaco-epidemiological studies, study of rare diseases and informing policy (2,4,5).

Despite the potential of primary care EHRs to facilitate a variety of secondary purposes, there are challenges with the use of this data source. Data within primary care EHRs are collected for clinical purposes and may therefore be influenced by the processes used to collect, amalgamate, extract and disseminate the data, which could lead to biased research outcomes (5,6). For instance, patients who attend general practice frequently are more likely to have complete records, as there is more opportunity for data to be captured by the general practitioner and recorded in the EHR (1). However, these patients may tend to be sicker and/or have better access to healthcare. Moreover, jurisdictions where certain chronic conditions are incentivised by governments may have more consistently recorded information in primary care EHRs compared with unincentivised conditions (6). These examples demonstrate that missing patient EHRs or clinical information from EHR databases may affect research validity or generalization of findings.

The lack of a standardized primary care electronic health record software system in both Australia and Canada presents further challenges for researchers intending on using these data. Inconsistent data formats and structures require complex data extraction and amalgamation processes which have the potential to introduce bias in the data (6,7). Further, practices utilising EHR software systems that allow for more structured data entry as opposed to free text fields may have more usable data given the former can be easily coded (5). Utilising EHR data from a subset of practices for research may provide a biased view of the disease under study.

Various countries have introduced incentive schemes to improve the quality of data recorded in EHRs (8–11). Whilst these programs demonstrated improved EHR data accuracy and completeness, and better standardization of recording practices, they also reported challenges in integration with workflow, unrealistic performance metrics and a focus on meeting numeric targets as opposed to ensuring data accuracy (8–11).

To minimise the likelihood of bias findings in research utilising EHRs, it has been recommended to first assess the quality of EHR data in the context in which the data will be used (12,13). The aim is to determine whether these data are suitable for the intended research purpose and to only proceed using these data for research once ‘fitness for use’ has been established (12,13). Various domains for assessing data quality have been proposed: completeness, plausibility, conformance, accuracy, currency and external validity (12–14). To date, very few studies utilising primary care EHRs have published data quality assessments beforehand (15,16). In recent years, automated data quality assessment programs have gained traction due to their efficiency and scalability (17–19). However, concerns remain around the accuracy and reliability of these tools, the lack of a standardised data quality assessment framework, and the ability to assess data quality in the context in which the data will be used (17–19).

Our research team has previously conducted a data quality assessment of Australian primary care EHR data in the context of osteoarthritis and developed a clinical prediction model for total knee replacement surgery from these data (16,20). Given the similarities in prevalence and management of osteoarthritis in Australia and Canada (21,22), our prediction model may be applicable to Canadian patients with osteoarthritis. This study aims to provide an overview of the quality of data recorded in Australian and Canadian primary care EHRs relating to osteoarthritis for the purposes of externally validating a clinical prediction model for total knee replacement surgery (16,20).

Osteoarthritis (OA) is a degenerative joint disease characterised by the breakdown of cartilage between bones (22,23). It affects 9% of Australians (approximately 2.2 million) and 14% of Canadians (approximately 3.9 million) (22,23). OA causes pain, inflammation, physical limitations and can have a significant impact on a person’s quality of life (23). OA has been ranked as the 13th leading cause of Years Lived with Disability (YLD) globally (24). Given OA is typically diagnosed and managed in community-based settings, primary care EHR data is an ideal source for OA research and surveillance.

Methods

Data sources

Australian data source

National Prescribing Service (NPS) MedicineWise is an Australian not-for-profit organisation focusing on the improvement of health through the appropriate use of medications and health technologies (25). They manage the MedicineInsight program which aims to identify key areas for improvement in primary care using data from consenting general practices across Australia (26). The MedicineInsight data set contains de-identified EHRs from over 2.9 million patients from 671 consenting general practices around Australia (26,27). Third party data extraction tools are used to de-identify, extract and securely transmit patient data monthly from general practice clinics to a central data warehouse (26). Here, structured data from two different EHR software systems, Medical Director and Best Practice (28,29) are merged into a consistent format before being provided to the researcher (26,30). The MedicineInsight data includes patient demographics, medications, diagnoses, procedures, clinical observations, pathology, allergy and alcohol status. Data fields containing raw text are provided as is, unless otherwise requested by the researcher. Medical Director uses the Docle diagnosis coding system and Best Practice uses Pyefinch (26). Clinicians do not always need to use a coded diagnosis and can enter diagnoses as free-text. Patient progress notes are not provided to researchers due to this field potentially containing patient identifying information.

Canadian data source

The Canadian Primary Care Sentinel Surveillance Network (CPCSSN) is a national collaboration of practice-based research networks (PBRN) collecting de-identified EHR data from contributing primary care physicians and nurse practitioners (31,32). The data are extracted, cleaned and processed bi-annually by each regional PBRN, then merged in a central repository located at Queen’s University in Kingston, Ontario, Canada. The 2018 national CPCSSN database included clinical information from nearly 1.8 million patients and over 1200 providers from 251 practices (31). The CPCSSN data contain the majority of patient information from the EHR, including demographics, diagnoses (current and historic), prescribed medications, physical examinations (height, weight, body mass index, blood pressure), laboratory results, referrals, risk factors, vaccinations and allergies (32). In Canadian primary care settings, the International Classification of Disease version 9 (ICD-9) is the standard system used for coding diagnoses and billing claims, though free text words also are used throughout the EHR (33).

Study sample

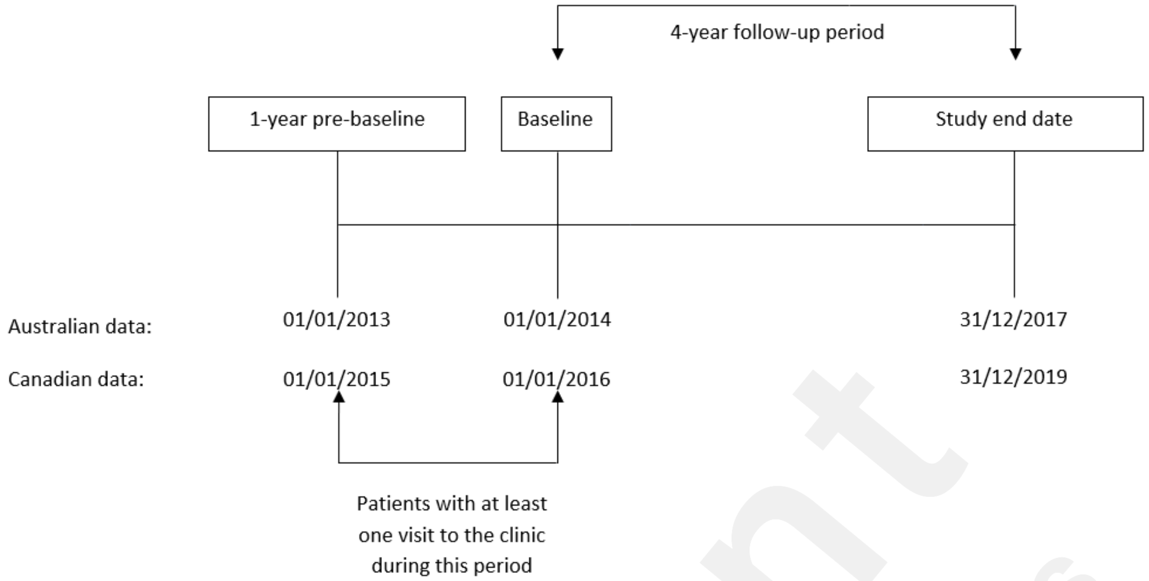
Australian study sample

NPS MedicineWise provided data extracted from 475,870 patient EHRs with a recorded diagnosis of osteoarthritis from 483 Australian general practice clinics. The coding used by NPS MedicineWise to identify patients with osteoarthritis has been provided in Multimedia Appendix 1. Patient socio-demographic and clinical data recorded in the EHRs were extracted up until the 31st of December 2017 (inclusive) and encounter data between the years 2013-2017 (inclusive). Clinical data included clinical observations, medications prescribed, pathology results, diagnoses and medical procedures.
Study baseline was defined as the 1st of January 2014. Patients were included in the study if they had attended general practice at least once in the year prior to baseline (2013) and were aged 45 years and over.

Canadian study sample

Patient data was extracted up until the 31st of December 2019 inclusive. The CPCSSN definition for osteoarthritis was used to identify patients with osteoarthritis (34). The definition uses ICD-9 codes (715, 721) found in the Billing table or the Problem List/Profile table (at least one at any time) to index a patient with osteoarthritis. When validated against medical chart review as the reference standard, the CPCSSN case definition showed reasonable accuracy with a sensitivity of 77.8%, specificity 94.9%, positive predictive value 87.7% and negative predictive value 90.2% (34). Multimedia Appendix 2 provides the definitions for osteoarthritis within the CPCSSN database. Study baseline was defined as the 1st of January 2016 for the Canadian cohort. Patients who were 45 years or older with at least one encounter with a primary care provider in the year prior to baseline (2015) were included in this study. Data from the encounter, billing, medications or diagnosis EHR tables were used to identify patient encounters. Patients assigned an “inactive” EHR status by their individual clinic or a “deceased” status with a date prior to baseline (2015) were removed from the analysis. Figure 1 shows the study timelines for the Australian and Canadian data sets.

Figure 1. Study timeline for Australian and Canadian data sets



Coding of variables

Coding of variables in Australian data

All socio-demographics, clinical observations and comorbidities were coded from the Australian EHRs at the 1st of January 2014. Geographical location of the patient and general practice was categorised using the Australian Bureau of Statistics (ABS) Australian Statistical Geography Standard (ASGS) remoteness areas (35): major cities of Australia, inner regional Australia, outer regional Australia, remote Australia and very remote Australia (35). These categories were dichotomised into urban and rural with remote and very remote Australia defined as rural. The Australian Index of Relative Socioeconomic Advantage and Disadvantage was coded by NPS MedicineWise using patient postcodes. These data were provided in quintiles with 1 representing those most disadvantaged and 5 those most advantaged.

The prevalence of commonly co-occurring chronic conditions in patients with osteoarthritis were summarised (23). These included hypertension, lipid disorder, ischaemic heard disease (IHD), depression, anxiety, asthma, diabetes mellitus, chronic obstructive pulmonary disease (COPD) and metastatic solid tumour. The prevalence of chronic conditions listed in the Charlson Comorbidity Index (CCI) were also summarised (36). The diagnosis onset date was used to determine which patients had a recorded diagnosis of these conditions in their EHR at study baseline. The coding used to identify patients with these chronic conditions is listed in Multimedia Appendix 1. For clinical observations, the latest weight, height and body mass index (BMI) recorded in the EHRs in the year prior to baseline was extracted.

Osteoarthritis medications were defined as medications belonging to the following ATC classes: H02 corticosteroids for systemic use, M01 anti-inflammatory and antirheumatic products, M02 topical products for joint and muscular pain, M09 other drugs for disorders of the musculo-skeletal system, N01 anaesthetics, N02 analgesics and N06 psychoanaleptics. The strength, dosage and frequency fields of the prescriptions data were used to calculate whether patients were likely to be taking osteoarthritis medications in the year prior to study baseline using prescriptions issued in the 12 months prior to baseline. Patients with prescriptions with missing strengths, dosages or frequencies were coded as having missing data for that medication as it was not possible to determine whether the patient was likely to be taking that medication at study baseline. Multimedia Appendix 1 contains a full list of the osteoarthritis medications included in this study.

Coding of variables in Canadian data

All socio-demographics, clinical observations and comorbidities were included in the analysis from the Canadian EHRs as of January 1st, 2016. Coding of variables in the Canadian EHR data was similar to the Australian data, except for geographical location of the patient, socio-economic status and prescribing of OA medications. Here, slight differences exist in the coding of variables due to differences in how these data are recorded in the Canadian and Australian EHR systems. The Canadian postal code was used to determine the geographical location of the patient. Postal codes that include a “0” as the second character were coded as rural delivery areas and the remaining as urban (37). Information from the Canadian Population Census was used to derive indicators for deprivation; residential instability, economic dependency, ethno-cultural composition and situational vulnerability, at a dissemination area level in the Canadian Index of Multiple Deprivation. The Postal Code Conversion File Plus (38) was used to assign a dissemination area and thus a quintile ranking of deprivation for each of the four dimensions of the CIMD to patients based on their recorded postcode in the EHR. The index was averaged across these four dimensions to get a composite index for socio-economic status where 1 represented least deprived and 5 most deprived.

Prescriptions were considered to be in use at study baseline if a patient had at least one record of a given osteoarthritis medication with a prescription start date within the pre-baseline period. EHR records with a missing start date of prescription were considered as having missing data. Multimedia Appendix 2 contain further information on how Canadian variables were coded for this analysis.

Data quality assessment

The quality of data in Australian and Canadian primary care EHRs was assessed using the data quality framework proposed by Kahn et al. (14). Data fields selected for assessment were those related to predictors from the clinical prediction model (20) and variables relevant to OA as determined through a literature search and an adapted Delphi process involving experts in the field of OA (39). Three domains of data quality were assessed where possible: (i) incompleteness, (ii) implausibility and (iii) external validity.

Incomplete and implausible data

Incompleteness was assessed by considering the context in which the data were collected and the management of OA in general practice. For example, weight is typically entered in the clinical observations field in Australian general practice EHRs and in an exam record in Canadian general practice EHRs. Hence, the proportion of patients without a recorded weight observation (Australian data) or weight exam record (Canadian data) in the year prior to baseline were summarised in addition to those with incomplete data in their clinical observation/exam record. Definitions for implausible data entries are listed in Table 1. Incomplete and implausible data were summarised using counts and percentages.

External validity

External validity of the prevalence of OA in the Canadian EHRs by age and gender were summarised and compared to national data from the Canadian Chronic Disease Surveillance System 2018 (40). Estimates were standardised by age and gender using the Canadian 2021 Census population as the reference standard (41). Equivalent comparisons were unable to be conducted in the Australian EHRs due to limited access to EHRs from patients without a diagnosis of OA. Instead, the recording of OA in the Australian EHRs was summarised by age and gender and estimates compared with data from the Australian Bureau of Statistics 2014-15 National Health Survey (42). The proportions from the National Health Survey were adjusted to account for the survey sampling strategy (43).

The recorded rates of total knee replacement surgery in the EHRs in the 4-year follow-up period were compared with national estimates from the Australian Institute of Health and Welfare National Hospital Morbidity Database (21) and the Canadian Joint Replacement Registry (44). Rates of total knee replacement surgery were presented per 100,000 people.

Table 1. Implausible data entry definitions

	Definition of implausible data entry	
Variable	**Australian EHR data**	**Canadian EHR data**
Age at study start	Year of birth beyond data extraction end date of 31st December 2017	Year of birth beyond data extraction end date of 31st December 2019
Height (cm)	Less than 100cm or greater than 240cm	Less than 120cm or greater than 200cm
Weight (kg)	Less than 20kg or greater than 280kg	Less than 20kg or greater than 280kg
BMI (kg/m^2)	Less than 12 kg/m^2 or greater than 90 kg/m^2	Less than 12kg/m^2 or greater than 90kg/m^2
Previous/contralateral TKR Any past knee surgery Total knee replacement	Before 01/01/1960 or year of surgery was before year of birth	
OA Medications	Prescription date before 01/01/1990	
Death	Before year of birth	

Results

Study cohorts

Selection of the Australian general practice EHR study cohort has been described elsewhere (16,20). In brief, 475,870 patient EHRs with a recorded diagnosis of OA were identified from the Australian MedicineInsight data set. Of these, 236,412 (49.7%) patient EHRs had a general practice encounter in the year prior to study baseline. A further 34,950 (7.3%) patient EHRs with an encounter recorded in the year prior to baseline were excluded: n=28,069 (5.9%) were less than 45 years old, n=2,117 (0.4%) had died prior to baseline and n=4,764 (1.0%) underwent bilateral TKR prior to study baseline. A total of 201,462 (42.3%) Australian general practice EHRs were available for analysis.

From the CPCSSN EHR database, 123,741 patient EHRs with a recorded diagnosis of OA were identified. A further 10,141 (8.2%) patient EHRs were excluded due to their inactive status, 19,631 (15.9%) patients did not have a record of attending general practice in the year prior to baseline, 1,538 (1.2%) patients were younger than 45 years at baseline and 6 (0.005%) had died prior to baseline. This resulted in a total of 92,425 (74.7%) Canadian general practice patient EHRs for data quality assessment.

Incomplete/missing data

The frequency and proportion of incomplete data fields are summarised in Table 2. There was minimal missing data for age, gender and geographical location in both the Australian and Canadian

EHRs. Just over half (52.9%) of the Canadian study population had a missing Canadian Index of Multiple Deprivation due to missing or incomplete postal code information. There were higher proportions of patients with missing dates of diagnosis for hypertension (9.5% vs 0.04%), lipid disorder (7.0% vs 0.08%), depression (9.6% vs 0.03%) and anxiety (17.7% vs 3.0%) in the Australian EHRs compared with the Canadian EHRs. The proportion of patients without an observation/exam record for height (64.3% vs 57.2%), weight (55.7% vs 58.1%) and BMI (68.0% vs 55.0%) in the year prior to study baseline was substantial in the two cohorts, and slightly higher for height and BMI in the Australian EHR data compared with the Canadian EHR data. Less than 1% of patients in the Canadian data set had a procedure record relating to knee surgery (including total knee replacement surgery) prior to and during the 4-year study period. This is in comparison to approximately 10% in the Australian cohort. There were higher proportions of missing data in OA medication records in the Australian EHR data compared with the Canadian data. However, different approaches were used to classify missing data in OA medication records between the data sets. In the Australian data set, missing OA medication data could be due to missing medication dosage, strength or frequency. These fields were not available in the Canadian data set and therefore only missing prescription dates were considered.

Missing/incomplete data

Table 2. Missing/incomplete data in Australian and Canadian primary EHRs

	Australian EHRs (MedicineInsight data) N=201,462		**Canadian EHRs (CPCSSN data) N=92,425**	
Patient characteristics	**n (%)**	**Missing/ incomplete data n (%)**	**n (%)**	**Missing/incomplete data n (%)**
Demographics				
Age (years), *mean (SD)*	67.2 (11.1)	3 (0.001)	67.6 (11.4)	-
Age categories (years)				
45-49	11,477 (5.7)		4,293 (4.6)	
50-64	72,300 (35.9)		34,462 (37.3)	
65-79	84,152 (41.8)		37,926 (41.0)	
80 years and over	33,530 (16.6)		15,744 (17.0)	
Gender		-		13 (0.01)
Female	123,376 (61.2)		57,157 (61.8)	

Male	78,049 (38.7)		35,255 (38.1)	
Other*	37 (0.02)		-	
Geographical location		1,068 (0.5)	-	3,845 (4.2)
Urban	173,296 (86.5)		70,023 (75.8)	
Rural	27,098 (13.5)		18,557 (20.1)	
Socio-economic status				
Australian Index of relative socio-economic advantage and disadvantage		1,228 (0.6)#		
1 (most disadvantaged)	41,600 (20.8)		N/A	
2	39,539 (19.8)			
3	48,122 (24.0)			
4 and 5 (most advantaged)	70,973 (35.5)			
Canadian Index of Multiple Deprivation**				48,854 (52.9)
1 (least deprived)			628 (1.4)	
2	N/A		15,750 (36.1)	
3			15,666 (36.0)	
4			11,210 (25.7)	
5 (most deprived)			317 (0.7)	
Comorbidities	**n (%)**	**Missing date of diagnosis n (%)**	**n (%)**	**Missing date of diagnosis n (%)**
Hypertension	81,004 (44.4)	19,037 (9.5)	38,944 (42.1)	36 (0.04)
Lipid disorder	58,478 (31.2)	14,192 (7.0)	54,922 (59.4)	76 (0.08)
Ischemic heart disease	23,522 (11.8)	2,810 (1.4)	10,272 (11.1)	2,603 (2.8)
Depression	35,644 (19.6)	19,393 (9.6)	22,337 (24.2)	25 (0.03)
Anxiety	16,373 (10.1)	35,644 (17.7)	18,219 (19.7)	2,771 (3.0)
Asthma	21,757 (11.2)	7,088 (3.5)	8,715 (9.4)	2,421 (2.6)
Diabetes mellitus	27,821 (14.1)	4,205 (2.1)	13,768 (14.9)	75 (0.08)
Chronic obstructive pulmonary disease	12,709 (6.4)	2,977 (1.5)	7,410 (8.0)	15 (0.02)
Metastatic solid tumour	33,485 (16.9)	3,470 (1.7)	334 (0.4)	108 (0.1)
	n (%)	**Missing date of diagnosis for at least one Charlson comorbidity**	**n (%)**	**Missing date of diagnosis for at least one Charlson comorbidity n (%)**

			n (%)			
Count of chronic conditions from the Charlson Comorbidity Index (CCI), *median [IQR]*	0 [0,1]		15,875 (7.9)	0 [0,1]		1,370 (1.5)
No conditions	113,097 (60.9)			42,087 (45.5)		
1 condition	49,216 (26.5)			26,510 (28.7)		
2 conditions	16,407 (8.8)			9,777 (10.6)		
3 or more conditions	6,267 (3.4)			4,637 (5.0)		
Clinical observations	**Mean (SD)**	**Missing data in observation record** **n (%)**	**Missing observation record** **n (%)**	**Mean (SD)**	**Missing a measurement in exam record** **n (%)**	**Missing exam record** **n (%)**
Height (cm)	165.3 (9.9)	-	129,552 (64.3)	165.7 (9.9)	1,254 (1.4)	51,613 (55.8)
Weight (kg)	82.1 (20.1)	-	112,230 (55.7)	84.6 (24.6)	9,007 (9.7)	44,647 (48.3)
BMI (kg/m^2)	30.1 (6.5)	-	137,071 (68.0)	30.5 (7.4)	270 (0.3)	50,533 (54.7)
Knee surgery	**n (%)**	**Date missing from knee procedure entry in diagnosis record** **n (%)**	**Missing knee procedure entry in diagnosis record** **n (%)**	**n (%)**	**Date missing from procedure record** **n (%)**	**Missing TKR procedure record** **n (%)**
TKR		1,254 (0.6)	182,128 (90.4)		121 (0.1)	92,261 (99.8)
TKR prior to study	9,432 (4.7)			16 (0.02)		
TKR during 4-year study period***	8,638 (4.3)			27 (0.03)		
Past knee surgery (excluding TKR)	6,070 (3.0)	990 (0.5)	194,385 (96.5)	76 (0.08)	28 (0.03)	92,252 (99.8)
OA medications prescribed	**n (%)**	**Missing data in medication$^{\lambda}$ record**	**Missing medication record** **n (%)**	**n (%)**	**Missing data in medication$^{\beta}$ record** **n (%)**	**Missing medication record** **n (%)**

		n (%)				
Prescribed ≥1 OA medication in last year	57,090 (33.8)	32,548 (16.2)	-	44,289 (41.9)	51 (0.06)	715 (0.8)
OA medications prescribed in last year:						
H02 Corticosteroids for systemic use	3,457 (1.9)	22,188 (11.0)		5,348 (5.8)		
M01 Anti-inflammatory & antirheumatic	21,220 (10.9)	5,834 (2.9)		17,500 (18.9)		
M02 Topical for joint & muscular pain	-	387 (0.2)		5,412 (5.9)		
M09 Other drugs for disorders of musculoskeletal system	-	58 (0.03)		-		
N01 Anesthetics	-	34 (0.02)		111 (0.1)		
N02 Analgesics	39,882 (22.3)	22,227 (11.0)		20,660 (22.4)		
N06 Psychoanaleptics	1,987 (1.0)	173 (0.09)		18,998 (20.6)		

	n (%)	**Missing date of death n (%)**	**n (%)**	**Missing date of death n (%)**
Death during 4-year study period	7,720 (3.9)	1,861 (0.9)	1,436 (1.6)	2,888 (3.1)

*Other gender category is not recorded in Canadian EHR data.

**Missing data included missing postal code information in CPCSSN (n=3,845) and incomplete information arising due to CPCSSN data with only first three digits of the postal code (n=42,300) and due to lack of assigned deprivation quantities for a subset of dissemination areas when a complete CPCSSN postal code was available (n=2,709).

*** Total knee replacements occurring between 2014-2017 for Australian data and between 2016-2019 for Canadian data.

#Unknown whether these represent patients without a postcode or whether these patients live in low population areas that do not have an allocated socio-economic status.

λMissing data included absence of medication dosage, strength and/or frequency.

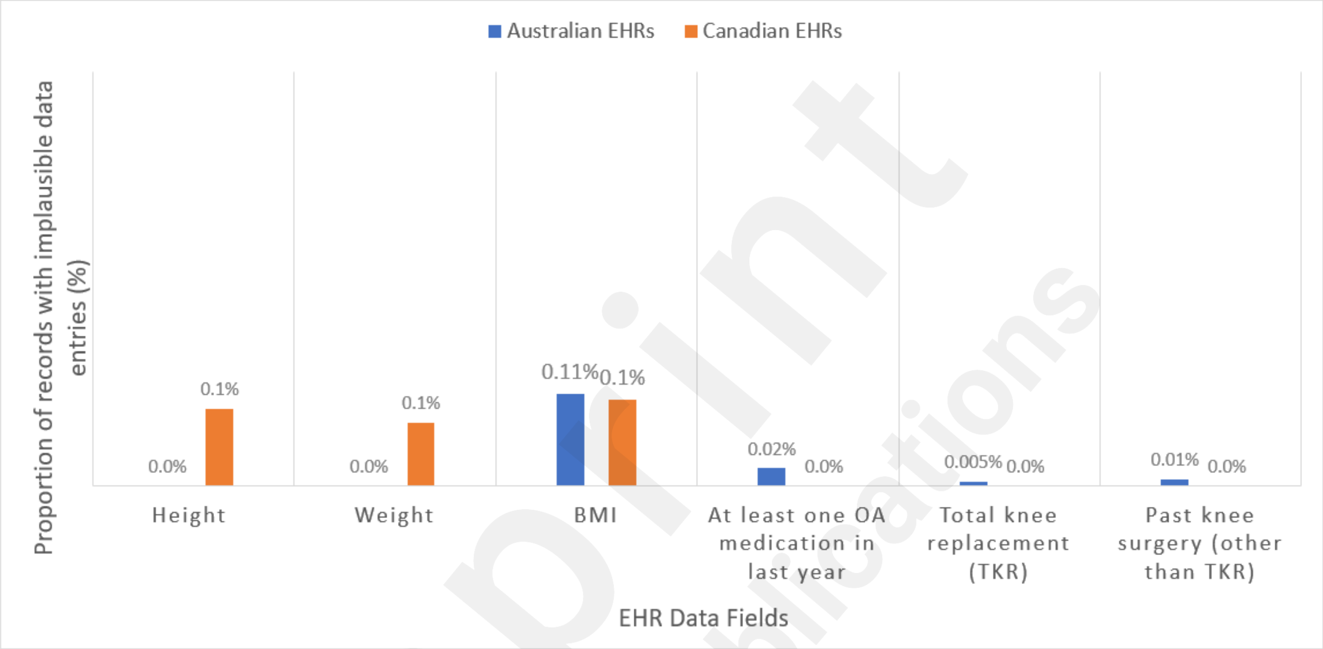
βMissing data included absence of medication code or a start date.

Abbreviations: EHR=electronic health record; CPCSSN=Canadian Primary Care Sentinel Surveillance Network; SD=standard deviation; CCI=Charlson Comorbidity Index; IQR=interquartile range; TKR=total knee replacement; OA=osteoarthritis; ATC=Anatomical Therapeutic Code.

Implausible data

Overall, there were minimal implausible data entries (Figure 2) for the data fields under study.

Figure 2. Proportion (%) of primary care EHRs with implausible data entries: Australia vs. Canada



External validity

Osteoarthritis by age and gender (Table 3) in the Australian EHR data were comparable with national estimates, except for females aged 80 years and over, where proportions were slightly higher in the EHR data compared with the National Health Survey. Age and sex standardised prevalence of osteoarthritis recorded in the Canadian EHRs were comparable with estimates from the Canadian Chronic Disease Surveillance System, except for the 80 years and over age group where prevalence was lower in the EHRs (36.7%, 95%CI 36.1 to 37.2 vs. 50.8%, 95% CI 50.7 to 50.9).

Total knee replacement rates recorded in the Australian and Canadian EHRs over the 4-year study period (Figure 3) were lower than their respective national estimates (Australian data: 72 per 100,000 EHR vs 218 per 100,000 national estimates; Canadian data: 0.84 per 100,000 EHR vs 200 per 100,000 national estimates).

Table 3. External validity of recording of OA in Australian and Canadian primary care EHRs

	OA (%) by age and gender in Australian EHR data **N=201,462**			**OA (%) by age and gender in Australia National Health Survey (2014-15)*** **N=1,933,849**		
Age	**Female**	**Male**	**Overall**	**Female**	**Male**	**Overa**

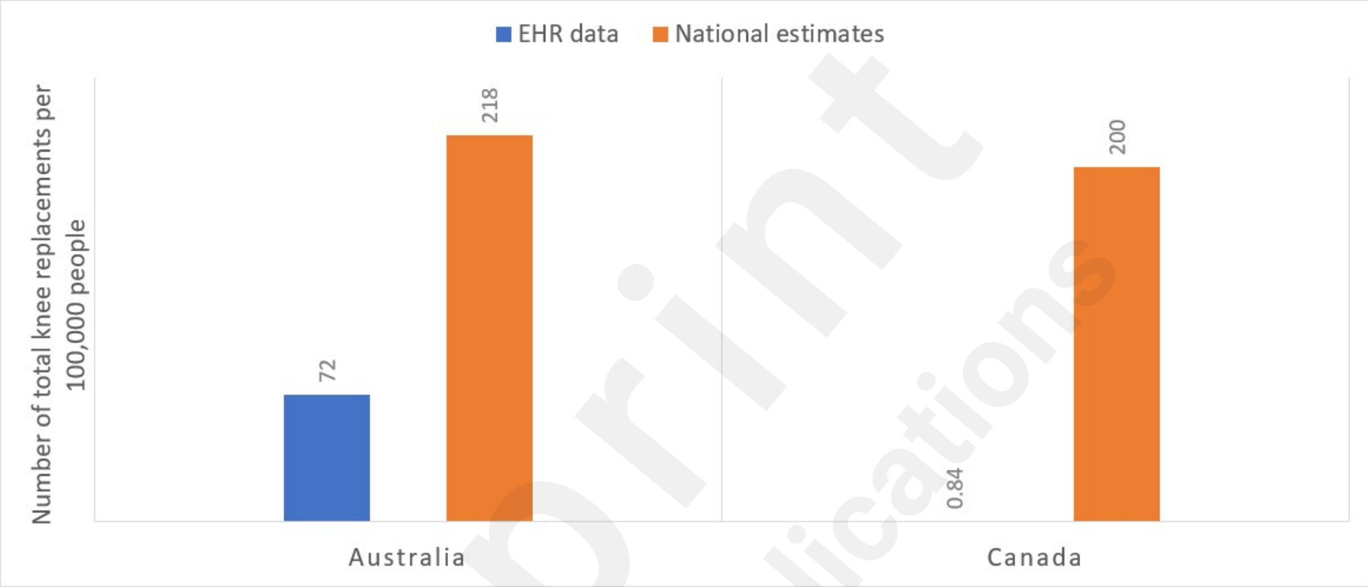
	% (95% CI)	% (95% CI)	% (95% CI)	% (95% CI)	% (95% CI)	% (95%
45-49	5.8 (5.5, 6.1)	5.5 (5.2, 5.8)	5.7 (5.4, 6.0)	5.3 (3.9, 7.2)	6.3 (4.2, 9.3)	5.6 (4.4, 7
50-64	35.7 (34.8, 36.5)	36.2 (35.4, 37.1)	35.9 (35.1, 36.7)	38.9 (35.9, 42.0)	42.7 (37.4, 48.2)	40.2 (37.7 42.8)
65-79	40.8 (40.1, 41.5)	43.3 (42.6, 44.0)	41.8 (41.1, 42.4)	41.8 (39.1, 44.6)	39.3 (34.5, 44.4)	41.0 (38.5 43.4)
80+	17.7 (17.0, 18.4)	15.0 (14.3, 15.6)	16.6 (16.0, 17.3)	13.9 (11.5, 16.8)	11.6 (8.7, 15.3)	13.1 (11.2 15.4)
	Prevalence (%) of OA in Canadian EHR data N=433,474			**Prevalence (%) of OA in Canadian Chro Disease Surveillance System (2018)** N=21,480,750**		
45-49€	7.7 (7.4, 8.0)	6.8 (6.5, 7.2)	7.3 (7.1 to 7.5)	4.2 (4.2, 4.2)	3.7 (3.6, 3.7)	3.9 (3.9, 4
50-64	18.5 (18.3, 18.8)	15.1 (14.9, 15.4)	17.0 (16.8 to 17.2)	17.8 (17.8, 17.8)	13.4 (13.3, 13.4)	15.6 (1 15.6)
65-79	33.0 (32.5, 33.4)	24.9 (24.5, 25.3)	29.3 (29.0 to 29.6)	38.3 (38.3, 38.4)	27.6 (27.6, 27.7)	33.2 (3 33.2)
80+	39.9 (39.1, 40.7)	31.7 (30.9, 32.6)	36.7 (36.1 to 37.2)	55.9 (55.7, 56.0)	43.2 (43.1, 43.4)	50.8 (5 50.9)
Overallβ	25.0 (24.8, 25.2)	19.0 (18.8, 19.2)	22.1 (22.0 to 22.3)	28.5 (28.5, 28.5)	20.0 (19.9, 20.0)	24.5 (2 24.5)

*From Australian Bureau of Statistics National Health Survey (NHS) 2014-15 (42)

**From Canadian Chronic Disease Surveillance System 2016(40).

€OA prevalence for 45-49 age category for Canadian Chronic Disease Surveillance System also includes data for those who are 35-44.
βAge and sex standardised ratios using the Canadian 2021 Census population as the reference standard.

Figure 3. External validity of recording of total knee replacement surgery in Australian and Canadian primary care EHRs (number of total knee replacements per 100,000 people)



*National estimates from Australian Institute of Health and Welfare National Hospital Morbidity Database (13) and the Canadian Joint Replacement Registry (35).

Comparing the characteristics recorded in general practice EHRs in patients with OA from Australia and Canada

Socio-demographic characteristics were similar between the Australian and Canadian cohorts, except a higher proportion of patients in the Australian cohort were from urban areas (86.5% vs 75.8%). Socio-economic status was not compared between the two countries due to high proportions of missing data for the Canadian Index of Multiple Deprivation and differences in the definitions of the measures. A lower proportion of patients had a recorded diagnosis of lipid disorder (31.2% vs 59.4%) and anxiety (10.1% vs 19.7%) in the Australian cohort compared with the Canadian cohort. A higher proportion of patients had a recorded diagnosis of metastatic solid tumour (16.9% vs 0.4%) in the Australian EHR data compared with the Canadian EHR data. Just over 60% of patients in the Australian EHR data set did not have a recorded diagnosis of a chronic condition listed in the Charlson Comorbidity Index compared with approximately 46% in the Canadian data set. Despite substantial amounts of missing data, weight, height and BMI were similar between the two cohorts, and the proportion of patients who died during the 4-year study follow-up was similar. A slightly higher proportion of patients (previous TKR: 4.7% vs 0.02%; TKR during study: 4.3% vs 0.03%; past knee surgery: 3.0% vs 0.08%) had knee surgery recorded in the Australian EHR data compared with the Canadian data, and a lower proportion of Australian patients (33.8% vs 41.9%) were prescribed at least one OA medication in the year prior to baseline. More specifically, lower proportions of prescribing of anti-inflammatory and antirheumatic products (10.9% vs 18.9%) and psychoanaleptics (1.0% vs 20.6%) were recorded in the Australian EHR data compared to the Canadian EHR data.

Discussion

Suitability of EHR data for clinical prediction model validation

This study assessed the quality of data related to OA recorded in Australian and Canadian general practice EHRs for the purposes of externally validating a clinical prediction model for knee replacement surgery. More specifically, the completeness, plausibility and external validity of data fields relating to osteoarthritis were assessed. Overall, the quality of data recorded in the Australian and Canadian EHRs were similar. Missing data were minimal for all sociodemographic characteristics of interest, except for socioeconomic status in the Canadian data (Canadian Index of Multiple Deprivation). Here missing data were due to patient EHRs containing only the first three digits of the postal code (n=42,300), missing postal code (n=3,845) or lack of assigned deprivation quantities for a subset of postal areas (n=2,709). While full postal codes are usually collected at the point of care, this is often considered identifiable patient information and is generally not permitted to be extracted and used for secondary data analysis; thus, the CPCSSN research database has a higher proportion of missing full postal code data and this field may not be useful for Canadian research studies as it currently exists. Further work is needed to understand whether missing Canadian Index of Multiple Deprivation data are likely to be missing completely at random (MCAR) or whether imputation methods for missing data that are missing at random (MAR) may be used to recover missing deprivation values (45,46).

Similarly, there were substantial missing data for weight, height and BMI in both the Australian and Canadian EHRs. BMI may be an indicator of OA disease progression, and therefore a potentially important data field for OA research (39). Results from this assessment suggest that weight, height and BMI extracted from EHRs from patients with OA may be unsuitable for research use due to large amounts of missing data. Utilising EHRs with complete BMI data only may lead to biased results if for example, patients who attend general practice more often tend to have more complete EHRs but tend to be in poorer health (1). In this scenario, utilising data from patients with complete EHR data only may lead to misrepresentation of the study population of interest as healthier patients are likely to be underrepresented in the study sample. Given Australian general practice guidelines suggest biennial BMI measurements (47) in all adults 18 years and over and Canadian guidelines every 1-3 years, further research is required to confirm whether BMI data may be more complete if extracted over a two-year (Australia) or three-year (Canada) period as opposed to one year pre-baseline. Further, some of the missing BMI values in both data sets may be due to the recording of this information in free text fields that are not currently extracted by CPCSSN or NPS MedicineWise (i.e. clinical progress notes). While there are suggested content standards for Canadian and Australian EHRs, these are not mandatory and thus, variation in data structures, formats and content continues to exist between the many EHR products available in both these countries.

Missing dates of diagnoses were less than 10% for comorbidities that commonly co-occur with OA, except for anxiety in the Australian EHR data. It is unknown exactly why close to 20% of records with an anxiety diagnosis in the Australian data had a missing diagnosis onset date. Some of this may be explained by patients having long standing anxiety, including undiagnosed anxiety, making it difficult for patients to recall the date of onset.

The prevalence of OA was similar in the Canadian EHRs compared with national estimates from the Canadian Chronic Disease Surveillance System (CCDSS), except for the 80+ age group where the prevalence was smaller in the EHRs. There were some differences in how OA cases were defined in each data source (multiple elements within the EHR versus hospital and billing records in the CCDSS), and the addition of hospital records in the CCDSS may have accounted for older patients with more recent diagnoses of OA that have not been documented in their general practice EHR but have undergone surgery.

Lastly, the rates of total knee replacement recorded in both the Australian and Canadian EHRs were markedly smaller than national estimates. In both countries, for a surgery such as total knee replacement to be recorded in a patient’s general practice EHR, the patient must either inform their general practitioner of the surgery and date or the general practice clinic receives documentation from the hospital notifying them of the surgery and date, and the general practitioner then enters this information into the EHR system. Typically, the documentation from the hospital is stored in the EHR but does not necessarily get entered as a procedure in the EHR which may explain the underrepresentation of total knee replacement surgery in this study cohort. Data relating to knee replacement surgery in Australian and Canadian general practice EHRs are likely to be unsuitable for use in OA research in its current state. Data linkage with national joint replacement registries or hospital databases may be required if researchers wish to conduct OA studies where the true rates of knee replacement in OA patients attending general practice is of interest.

Strengths and limitations

This study contributes to the limited literature on the quality of data in general practice EHRs. It is the first study, to our knowledge, to compare the quality of these data internationally and in the context of a globally important chronic disease, OA. This study provides insight into specific data fields in EHRs that can be targeted for more complete recording in general practice or potential data fields for the development and testing of novel missing data methods. Further, the data quality assessment methods used in this study were based on established data quality assessment guidelines (14). Study cohort sizes were large in both data sets and therefore likely to provide a true representation of data recorded in general practice EHRs for patients with OA. Lastly, this data quality assessment was conducted with input from epidemiologists, biostatisticians and general practitioners and considers the context in which these data were collected.

Due to limited national OA patient data sets available for data linkage in both countries, accuracy of the EHR data were unable to be assessed. Further, there were differences in the dates of data extraction between the Australian and Canadian EHR data sets, with the latter containing more recent patient EHR data. Whilst this may seem problematic at first, we applied a consistent study timeline (i.e., 4-year follow-up period) and it is unlikely that recording practices relating to OA in the Australian EHRs have changed significantly between 2013-2017 to 2015-2019. There were also slight differences in the coding of osteoarthritis prescriptions between the Australian and Canadian EHR data sets due to limitations in the prescriptions data fields available for analyses in the Canadian data set. This may explain the relatively higher amounts of missing data for OA prescriptions in the Australian data set where missing data in any of the medication strength, dosage or frequency fields would result in missing data for that particular OA medication. In the Canadian data set, missing OA prescriptions data arose from missing medication codes and missing associated start dates only.

Furthermore, data quality may have impacted the study population due to inclusion and exclusion criteria. A patient was included in the study population based on the condition of having at least one encounter within pre-baseline period. This could potentially artificially select for more complete EHR records and thus indicate a higher quality of data than what is available in the data sources that were under investigation for this study.

There were differences in the prescribing of anti-inflammatory and antirheumatic products (M01) and psychoanaleptics (N06) in the Australian and Canadian EHR data sets, with higher rates reported in the Canadian EHR data. In both countries, recording of over-the-counter medications in general practice EHRs requires the patient to recall this information, inform their general practitioner and the general practitioner to record this information in the designated area of the EHR. It is possible that these medications are not captured well in Australian EHRs or are being entered elsewhere in the EHR. Reasons for the low rates of recording of N06 medications in the Australian cohort remains unknown and warrants further investigation.

Due to limited national data available on patients with OA in both countries, we were unable to externally validate all EHR data fields of interest in this study. However, from previous work conducted by NPS MedicineWise (26) and our research team on the Australian EHR data set (16), external validity has been assessed for the recording of OA prevalence, remoteness areas, BMI, comorbidities (hypertension, lipid disorder, ischaemic heart disease, asthma, diabetes, chronic obstructive pulmonary disease, metastatic solid tumour, depression and anxiety) and prescribing of OA medications through comparison with the 2014-15 National Health Survey. The assessment demonstrated good external validity for these data fields except for the prescribing of OA medications (Australian EHRs 34% vs National estimate 55%) and metastatic solid tumour (Australian EHRs 17% vs National estimate 26%). The Australian National Health Survey asks participants to report all medication usage, including over the counter medications and medications prescribed by specialists. These may not be captured in general practice EHRs and may explain differences in OA medication rates between the Australian EHR data and Australian National Health Survey estimates. Further work is needed to externally validate the prescribing rates of OA medications reported in this study. The lower rates of recorded metastatic tumours in Australian general practice EHRs may be due to tumour diagnoses by specialists not being communicated to the general practitioner. Hence, data relating to metastatic tumours in the Australian EHRs may not be fit for use in research.

Lastly, whilst this study aimed to address the quality of data in Australian and Canadian general practice EHRs in the context of validating a clinical prediction model, the useability of these data for research is also worth noting. It takes a significant amount of time to adequately clean and prepare EHR data for analysis, including quality assessments (32,48). Text-heavy fields, such as prescriptions and diagnoses data, often contained typographical errors and required advanced pattern-matching searches to identify certain conditions or medications. Extensive data cleaning and pre-processing was conducted on both the Australian and Canadian data sets prior to assessing data quality in this study. Researchers wanting to utilise these data for research purposes should be made aware of the effort required to code and prepare general practice EHR data for research use. This work also highlights the strong need for better standardisation of general practice EHR software systems and development of natural language processing software specific to general practice EHR data.

Conclusions

This study assessed the quality of data in Australian and Canadian general practice EHRs in the context of OA for the purposes of validating a clinical prediction model. Overall, data quality was similar in the two data sets. Missing and implausible data were minimal except for the recording of weight, height, BMI and Canadian Index of Multiple Deprivation. These data fields may not be fit for use in OA research due to large proportions of missing data which are unlikely to be recoverable using imputation techniques. Further work is required to better understand the nature of these missing data. External validity of recording of knee surgery in both the Australian and Canadian EHRs was poor. More accurate recording of surgeries in general practice EHRs is required if these data are to be utilised in OA research. In the meantime, data linkage with national joint replacement and surgical registries may overcome some of these data quality challenges.

Acknowledgements

We would like to acknowledge NPS Medicinewise for providing the MedicineInsight primary care EHR data, the Australian Bureau of Statistics for the National Health Survey data and the Canadian Primary Care Sentinel Surveillance Network (CPCSSN). This work is supported by the National Health and Medical Research Council of Australia (NHMRC) Centre for Research Excellence in Total Joint Replacement (APP1116325). Sharmala Thuraisingam is the recipient of a scholarship awarded through the NHMRC Centre for Research Excellence in Total Joint Replacement (APP1116235). Jo-Anne Manski-Nankervis holds a Medical Research Future Fund Next Generation Clinical Researchers Program—Translating Research into Practice Fellowship (1168265). Michelle Dowsey holds a University of Melbourne Dame Kate Campbell Fellowship. We would like to acknowledge Ken Martin and Rachael Morkem for their assistance in obtaining the CPCSSN data required for this analysis.

Conflicts of Interest

The authors declare that they have no competing interests in relation to this study.

Abbreviations

ABS	Australian Bureau of Statistics
ASGS	Australian Standard Geographical
ATC	Anatomical Therapeutic Code
CCI	Charlson comorbidity index
CPCSSN	Canadian Primary Care Sentinel Surveillance Network

EHR	Electronic Health Record
ICD	International Classification of Diseases
IQR	Interquartile range
NPS	National Prescribing Service
OA	Osteoarthritis
PBRN	Practice-based research network
SD	Standard deviation
YLD	Years Lived with Disability

Data Availability

The data that support the findings of this study are available from NPS MedicineWise, Australian Bureau of Statistics and the Canadian Primary Care Sentinel Surveillance Network but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data may be available from the authors upon reasonable request if permission is granted by the parties listed above. The data sets provided by NPS MedicineWise and the Canadian Primary Care Sentinel Surveillance Network are not publicly available due to patient privacy.

Multimedia Appendix 1.

Coding for Australian EHR data (DOCX. File)

Multimedia Appendix 2.

Coding for Canadian EHR data (DOCX. File)

References

1. Wells BJ, Chagin KM, Nowacki AS, Kattan MW, Chagin KM, Kattan MW. Strategies for handling missing data in electronic health record derived data. EGEMS. 2013;1(3):1035.
2. Birtwhistle R, Williamson T. Primary care electronic medical records: a new data source for research in Canada. CMAJ Canadian Medical Association Journal. 2015;187(4):239–40.
3. Australian Institute of Health and Welfare. Frequent GP attenders and their use of health services in 2012–13, Summary - Australian Institute of Health and Welfare [Internet]. 2015 [cited 2023 Jan 25]. Available from: https://www.aihw.gov.au/reports/primary-health-care/frequent-gp-attenders-use-health-services-2012-13/contents/summary
4. de Lusignan S, Metsemakers JF, Houwink P, Gunnarsdottir V, van der Lei J. Routinely collected general practice data: goldmines for research? A report of the European Federation for Medical Informatics Primary Care Informatics Working Group (EFMI PCIWG) from MIE2006, Maastricht, The Netherlands. Informatics in Primary Care. 2006;14(3):203–9.
5. de Lusignan S, van Weel C. The use of routinely collected computer data for research in primary care: opportunities and challenges. Family Practice. 2006 Apr 1;23(2):253–63.
6. Verheij RA, Curcin V, Delaney BC, McGilchrist MM. Possible Sources of Bias in Primary Care

Electronic Health Record Data Use and Reuse. Journal of medical Internet research. 2018 May 29;20(5):e185.

7. Liaw ST, Taggart J, Yu H, Lusignan S de, Kuziemsky C, Hayen A. Integrating electronic health record information to support integrated care: Practical application of ontologies to improve the accuracy of diabetes disease registers. Journal of Biomedical Informatics. 2014 Dec 1;52:364–72.
8. Scott A, Sivey P, Ait Ouakrim D, Willenberg L, Naccarella L, Furler J, et al. The effect of financial incentives on the quality of health care provided by primary care physicians. The Cochrane database of systematic reviews. 2011 Sep 7;(9).
9. Blumenthal D, Tavenner M. The “Meaningful Use” Regulation for Electronic Health Records. New England Journal of Medicine. 2010 Aug 5;363(6):501–4.
10. Trout KE, Chen LW, Wilson FA, Tak HJ, Palm D. The Impact of Meaningful Use and Electronic Health Records on Hospital Patient Safety. International Journal of Environmental Research and Public Health. 2022 Oct 1;19(19):12525.
11. Policy Research Unit in the Commissioning and Healthcare System. Review of the Quality and Outcomes Framework in England. 2016.
12. Huang Y, Voorham J, Haaijer-Ruskamp FM. Using primary care electronic health record data for comparative effectiveness research: experience of data quality assessment and preprocessing in The Netherlands. Journal of Comparative Effectiveness Research. 2016;5(4):345–54.
13. Terry AL, Stewart M, Cejic S, Marshall JN, De Lusignan S, Chesworth BM, et al. A basic model for assessing primary health care electronic medical record data quality. BMC Medical Informatics and Decision Making. 2019 Feb 12;19(1):1–11.
14. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. eGEMs (Generating Evidence & Methods to improve patient outcomes). 2016 Sep 11;4(1):18.
15. Staff M, Roberts C, March L. The completeness of electronic medical record data for patients with Type 2 Diabetes in primary care and its implications for computer modelling of predicted clinical outcomes. Primary care diabetes. 2016 Oct 1;10(5):352–9.
16. Thuraisingam S, Chondros P, Dowsey MM, Spelman T, Garies S, Choong PF, et al. Assessing the suitability of general practice electronic health records for clinical prediction model development: a data quality assessment. BMC Medical Informatics and Decision Making 2021 21:1. 2021 Oct 30;21(1):1–11.
17. Ramakrishnaiah Y, Macesic N, Webb GI, Peleg AY, Tyagi S. EHR-QC: A streamlined pipeline for automated electronic health records standardisation and preprocessing to predict clinical outcomes. Journal of Biomedical Informatics. 2023 Nov 1;147:104509.
18. Ozonze O, Scott PJ, Hopgood AA. Automating Electronic Health Record Data Quality Assessment. Journal of Medical Systems. 2023 Dec 1;47(1):1–16.
19. Lewis AE, Weiskopf N, Abrams ZB, Foraker R, Lai AM, Payne PRO, et al. Electronic health record data quality assessment and tools: a systematic review. Journal of the American Medical Informatics Association : JAMIA. 2023 Sep 25;30(10):1730.
20. Thuraisingam S, Chondros P, Manski-Nankervis JA, Spelman T, Choong PF, Gunn J, et al. Developing and internally validating a prediction model for total knee replacement surgery in patients with osteoarthritis. Osteoarthr Cartil Open. 2022 Sep 1;4(3):100281.
21. Australian Institute of Health and Welfare. AIHW website. 2020 [cited 2018 Mar 9]. Osteoarthritis, Australian Institute of Health and Welfare. Available from:

https://www.aihw.gov.au/reports/arthritis-other-musculoskeletal-conditions/osteoarthritis/contents/what-is-osteoarthritis

22. Government of Canada. Osteoarthritis in Canada [Internet]. 2020 [cited 2021 Jul 27]. Available from: https://www.canada.ca/en/public-health/services/publications/diseases-conditions/osteoarthritis.html
23. Australian Institute of Health and Welfare. Osteoarthritis snapshot, What is osteoarthritis? - Australian Institute of Health and Welfare [Internet]. 2020 [cited 2019 Jan 14]. Available from: https://www.aihw.gov.au/reports/chronic-musculoskeletal-conditions/osteoarthritis/contents/what-is-osteoarthritis
24. World Health Organisation. Global health estimates: Leading causes of DALYs [Internet]. 2020 [cited 2021 Jul 27]. Available from: https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/global-health-estimates-leading-causes-of-dalys
25. NPS MedicineWise. About us - NPS MedicineWise [Internet]. 2020 [cited 2021 Jul 30]. Available from: https://www.nps.org.au/about-us
26. Busingye D, Gianacas C, Pollack A, Chidwick K, Merrifield A, Norman S, et al. Data Resource Profile: MedicineInsight, an Australian national primary health care database. International Journal of Epidemiology. 2019 Dec 1;48(6):1741–1741h.
27. MedicineWise N. MedicineInsight Data Book. 2018.
28. Best Practice Software Pty. Ltd. Best Practice Software – An Evolution In Practice Management [Internet]. 2020 [cited 2020 Oct 9]. Available from: https://bpsoftware.net/
29. Health Communication Network. Software Solutions for Medical Practitioners | MedicalDirector [Internet]. 2020 [cited 2020 Oct 16]. Available from: https://www.medicaldirector.com/
30. Daniels B, Havard A, Myton R, Lee C, Chidwick K. Evaluating the accuracy of data extracted from electronic health records into MedicineInsight, a national Australian general practice database. International Journal of Population Data Science. 2022 Jun 29;7(1).
31. Canadian Primary Care Sentinal Surveillance Network. Canadian Primary Care Sentinel Surveillance Network (CPCSSN) [Internet]. 2020 [cited 2021 Jul 30]. Available from: http://cpcssn.ca/
32. Garies S, Birtwhistle R, Drummond N, Queenan J, Williamson T. Data Resource Profile: National electronic medical record data from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN). International Journal of Epidemiology. 2017 Aug 1;46(4):1091–1092f.
33. World Health Organisation. WHO. World Health Organization; 2018 [cited 2018 Jun 7]. WHO | International Classification of Diseases. Available from: http://www.who.int/classifications/icd/en/
34. Williamson T, Green ME, Birtwhistle R, Khan S, Garies S, Wong ST, et al. Validating the 8 CPCSSN case definitions for chronic disease surveillance in a primary care database of electronic health records. Annals of Family Medicine. 2014;12(4):367–72.
35. Australian Bureau of Statistics. Australian Statistical Geography Standard (ASGS) [Internet]. 2021 [cited 2020 Apr 25]. Available from: https://www.abs.gov.au/websitedbs/D3310114.nsf/home/Australian+Statistical+Geography+Standard+(ASGS)
36. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. Journal of Chronic Diseases. 1987;40(5):373–83.
37. Statistics Canada. How Postal Codes Map to Geographic Areas: Discussion. 2007.
38. Statistics Canada. Postal Code OM Conversion File Plus (PCCF+) [Internet]. 2023 [cited 2023

Jan 25]. Available from: https://www150.statcan.gc.ca/n1/en/catalogue/82F0086X

39. Thuraisingam S, Dowsey M, Manski-Nankervis JA, Spelman T, Choong P, Gunn J, et al. Developing Prediction Models for Total Knee Replacement Surgery in Patients with Osteoarthritis: Statistical Analysis Plan. Osteoarthritis and Cartilage Open. 2020 Nov 24;100126.
40. Government of Canada. Canadian Chronic Disease Surveillance System (CCDSS) [Internet]. 2021 [cited 2022 Dec 19]. Available from: https://health-infobase.canada.ca/ccdss/data-tool/
41. Statistics Canada. Data tables, 2021 Census of Population [Internet]. 2022 [cited 2022 Dec 21]. Available from: https://www12.statcan.gc.ca/census-recensement/2021/dp-pd/dt-td/index-eng.cfm
42. Australian Bureau of Statistics. National Health Survey First Results. 2015.
43. Donath SM. How to calculate standard errors for population estimates based on Australian National Health Survey data. Australian and New Zealand Journal of Public Health. 2005 Dec;29(6):565–71.
44. Canadian Institute for Health Information, Canadian Joint Replacement Registry. Hip and Knee Replacements in Canada: CJRR Annual Statistics Summary, 2019–2020 | CIHI [Internet]. 2021 [cited 2021 Jul 27]. Available from: https://www.cihi.ca/en/hip-and-knee-replacements-in-canada-cjrr-annual-statistics-summary-2019-2020
45. Lee KJ, Roberts G, Doyle LW, Anderson PJ, Carlin JB. Multiple imputation for missing data in a longitudinal cohort study: a tutorial based on a detailed case study involving imputation of missing outcome data. International Journal of Social Research Methodology. 2016 Sep 2;19(5):575–91.
46. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009 Jun 29;338(7713):157–60.
47. Royal Australian College of General Practitioners. Guidelines for preventive activities in general practice 9th edition. 2016.
48. Garies S, Cummings M, Forst B, McBrien K, Soos B, Taylor M, et al. Achieving quality primary care data: a description of the Canadian Primary Care Sentinel Surveillance Network data capture, extraction, and processing in Alberta. International journal of population data science. 2019 Nov 20;4(2).

Supplementary Files

Multimedia Appendixes

Coding for Australian EHR data.
URL: http://asset.jmir.pub/assets/4e15e67ca0bf938c4218a50c17a7b07d.docx

Coding for Canadian EHR data.
URL: http://asset.jmir.pub/assets/5e37ac5476178c8a7c54fe2916a89f0d.docx