International Journal of Population Data Science

Journal Website: www.ijpds.org

Evaluating the accuracy of data extracted from electronic health records into MedicineInsight, a national Australian general practice database

Benjamin Daniels[1,2], Alys Havard[1,2,3], Rimma Myton[1], Cynthia Lee[1], and Kendal Chidwick[1,*]

Submission History

Submitted:	08/12/2021
Accepted:	06/04/2022
Published:	29/06/2022

[1]NPS MedicineWise, Level 7 / 418a Elizabeth St, Strawberry Hills, NSW, 2012, Sydney, Australia
[2]Medicines Policy Research Unit, Centre for Big Data Research in Health, UNSW Sydney, Australia
[3]National Drug and Alcohol Research Centre, UNSW Sydney, Australia

Abstract

Introduction
MedicineInsight is a database containing de-identified electronic health records (EHRs) from over 700 Australian general practices. Previous research validated algorithms used to derive medical condition flags in MedicineInsight, but the accuracy of data fields following EHR extractions from clinical practices and data warehouse transformation processes have not been formally validated.

Objectives
To examine the accuracy of the extraction and transformation of EHR fields for selected demographics, observations, diagnoses, prescriptions, and tests into MedicineInsight.

Methods
We benchmarked MedicineInsight values against those recorded in original EHRs. Forty-six general practices contributing data to MedicineInsight met our eligibility criteria, eight were randomly selected, and four agreed to participate. We randomly selected 200 patients $\geq$18 years of age within each participating practice from MedicineInsight. Trained staff reviewed the original EHRs for the selected patients and recorded data from the relevant fields. We calculated the percentage of agreement (POA) between MedicineInsight and EHR data for all fields; Cohen's Kappa for categorical and intra-class correlation (ICC) for continuous measures; and sensitivity, specificity, and positive and negative predictive values (PPV/NPV) for diagnoses.

Results
A total of 796 patients were included in our analysis. All demographic characteristics, observations, diagnoses, prescriptions and random pathology test results had excellent (>90%) POA, Kappa, and ICC. POA for most recent pathology/imaging test was moderate (81%, [95% CI: 78% to 84%]). Sensitivity, specificity, PPV, and NPV were excellent (>90%) for all but one of the examined diagnoses which had a poor PPV.

Conclusions
Overall, our study shows good agreement between the majority of MedicineInsight data and those from original EHRs, suggesting MedicineInsight data extraction and warehousing procedures accurately conserve the data in these key fields. Discrepancies between test data may have arisen due to how data from pathology, radiology and other imaging providers are stored in EHRs and MedicineInsight and this requires further investigation.

Keywords
primary care; validation; electronic health records

*Corresponding Author:
Email Address: KChidwick@nps.org.au (Kendal Chidwick)

https://doi.org/10.23889/ijpds.v7i1.1713

Introduction

Electronic health record (EHR) systems are widely used in primary care settings to maintain patients' clinical data and improve patient care [1, 2]. These systems typically record and store data related to patients' demographic characteristics, medical observations, diagnoses, medications prescribed and tests ordered/performed. Over the past two decades, the development of software tools to interface with these clinical information systems (CIS) and extract data from EHRs has resulted in the increased secondary use of EHR data for research purposes [3, 4]. To date, research utilising EHRs has come from data collections such as the UK's Clinical Practice Research Datalink (CPRD) [5] and THIN [6]; Canada's Electronic Medical Record Administrative Data Linked Database (EMRALD) [7] and Canadian Primary Care Sentinel Surveillance Network (CPCSSN) [8]; and the United States' Veterans Affairs Corporate Data Warehouse (CDW) [3, 9].

In Australia, despite widespread and early use of EHRs in primary care patient management [10], there has been limited use of EHRs for health research [11, 12]. Aggregating EHR data from multiple, different CISs, general practitioners' (GPs) and patients' concerns around secondary use, and a lack of incentives for practices to share their EHRs have all been cited as barriers to accessing these data for research purposes in Australia [11, 12]. Additionally, many practices use CISs with different coding systems, such as 'Docle', Pyefinch' and 'ICPC2+', and with different fields for collecting the same information, leading to a lack of consistency in the resulting data [13, 14]. In 2011, NPS MedicineWise launched the MedicineInsight program to navigate these challenges and build a secure, centralised national primary care data repository under a rigorous data governance framework [12, 13]. Over the past decade, these data have been used by academics and NPS MedicineWise to inform industry and government agencies about primary care and post-market surveillance of medicines, informing policy and supporting quality improvement activities in general practice [15]. There are numerous reports and over 40 peer reviewed publications to date based on MedicineInsight data including studies investigating Australian opioid use [16, 17], vaccination [18–20], and the management of chronic hepatitis C in general practice [21, 22].

MedicineInsight collects EHR data from over 700 Australian general practices (representing 9% of Australian GPs) [13, 23]. All MedicineInsight practices use either the Best Practice^{TM} (BP) or Medical Director^{TM} (MD) CIS (the CIS used by 81% of general practices in Australia) [24] and the GeneRic Health Network Information Technology for the Enterprise (GRHANITE^{TM}) [25] and INCA [26] data tools are used to extract data from the CIS. Before the EHRs are ready for research use de-identified data are: extracted using one of the extraction tools; transferred to a secure data warehouse where they undergo harmonisation and cleaning; processed further into a monthly snapshot to derive new variables, such as Anatomical Therapeutic Chemical (ATC) codes [27], clinical indicators and medical condition flags [28]; and finally stored as longitudinal data (Figure 1). Inaccuracies in the MedicineInsight data collection could potentially arise during any of these steps on the data's journey from CIS to the MedicineInsight database. Previous studies have highlighted problems with different EHR data extraction tools including proprietary, 'black box' coding and a lack of harmonised metadata between extraction tools and EHR systems [4, 29]. There is also the potential for errors to arise when harmonising fields across CIS or when deriving variables from existing fields with missing data.

Previous research validated the algorithms used to create medical conditions flags—indicators derived from various fields of a patient's EHR indicating whether the patient has a specific medical condition—in MedicineInsight [28] against gold-standard patient EHRs from participating general practices. However, the accuracy of the extracted and transformed data has not yet been formally validated.

The aim of our study was to measure how accurately the data from the CIS are processed into the MedicineInsight database after extraction from the practice CIS and transmission to the data warehouse. Specifically, we aimed to assess agreement for patient characteristics, patient medical observations, diagnoses, prescribed medicines and tests, and demographic data.

Methods

Study design

We conducted a validation study comparing MedicineInsight records to original EHRs from a sample of four general practice clinics located in the capital cities of Australia's two most populous states: Sydney, New South Wales and Melbourne, Victoria.

Study population

General practices

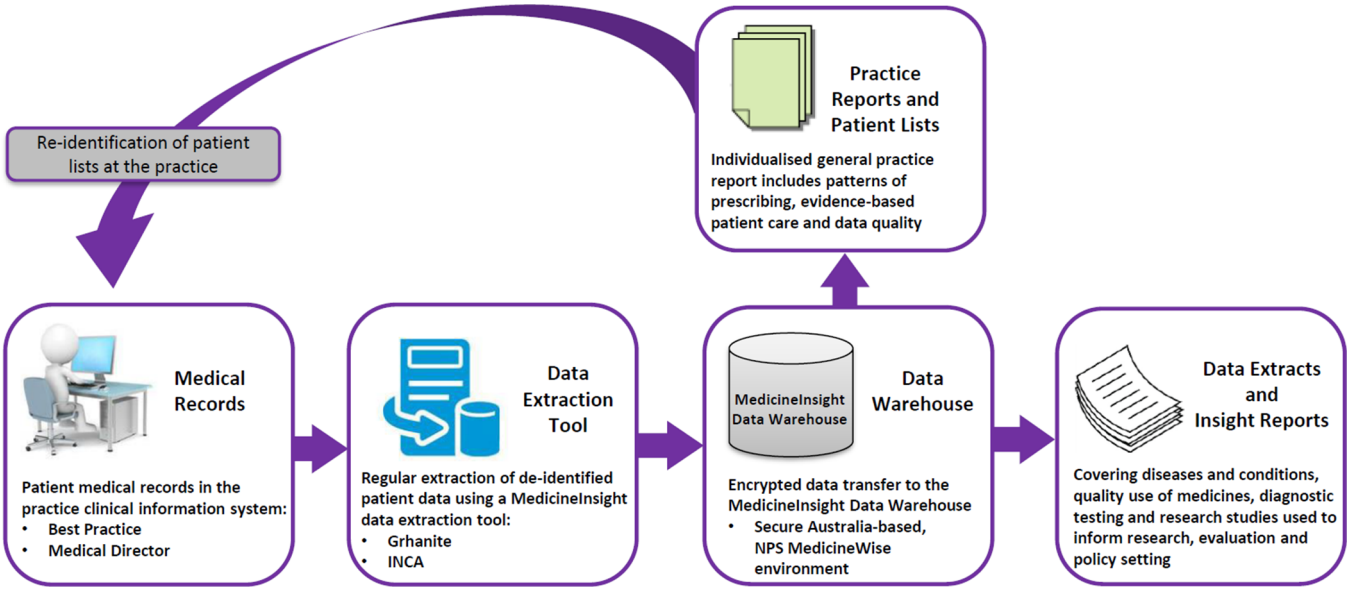
We identified 46 eligible practices participating in MedicineInsight from which eight were invited via stratified (by CIS and extraction tool) random selection and four agreed to participate—three located in Sydney and one in Melbourne. For a practice to be eligible for inclusion in our study, it had to:

- have at least 200 patients aged 18 years and older with at least three encounters in the two years prior to November 2020;
- be located within 40 kilometres of the Sydney or Melbourne central business districts;
- have participated in at least one MedicineInsight quality improvement activity in the period November 2019 to October 2020 to ensure a current interest in engaging with the MedicineInsight program; and
- have experience re-identifying patients in their CIS based on the IDs stored in MedicineInsight from previous quality improvement activities.

We stratified the random selection of practices according to the extraction tool and then the CIS used in the practice:

- GRHANITE^{TM} extraction tool and BP^{TM} CIS,

Figure 1: MedicineInsight electronic health record (EHR) data extraction, transformation, warehousing, and reporting procedures



- INCA^{TM} and BP^{TM} CIS,
- GRHANITE^{TM} extraction tool and MD^{TM} CIS,
- INCA^{TM} and MD^{TM} CIS.

Patient population

We randomly selected 200 patients from the MedicineInsight database for each of the four practices. Eligible patients were 18 years and older; had at least three encounters in the two years prior to November 2020; and had a CIS status of 'active' or 'deceased'. We excluded patients whose CIS status was 'inactive', 'visitor' or 'next of kin'.

Study period and data fields evaluated

Our study covered all activity at each general practice site between 1st November 2018 and 31st October 2020, unless otherwise specified. Similar to previous validation studies [7, 29], we assessed the accuracy of the data fields used to derive MedicineInsight condition flags, prescribing information, pathology tests, and patient demographics. Specifically, we examined:

1. The most recent patient observations, recorded at any time prior to 31st October 2020, for diastolic and systolic blood pressures, height, and weight.

2. Historical diagnoses recorded at any time prior to 31st October 2020 in the 'medical history' table. Due to time constraints only four diagnoses were chosen to be assessed, including two common chronic conditions, previously included in MedicineInsight quality improvement and research activities, (hypertension and osteoarthritis) and two serious acute conditions of interest for future MedicineInsight research activities (myocardial infarction, and venous thromboembolism (VTE)). Patients were flagged as having a recorded diagnosis only if they had a relevant coded or free text entry in the 'Diagnosis reason' field. Supplementary Appendix 1 includes the full list of terms used to define the conditions.

3. Most recently prescribed medicine within 24 months of 31st October 2020 as recorded in the 'Script item' table (name of medicine and date of prescription).

4. A prescribed medicine from a randomly selected, 'nominated' date from any point prior to 31st October 2020 as recorded in the 'Script item' table. We assessed the agreement of the medicine name, strength, quantity prescribed, number of repeats prescribed, and the Pharmaceutical Benefits Scheme (PBS) status for the prescription (PBS/PBS authority/Repatriation PBS (RPBS)/RPBS authority/private prescription) [30] for each prescription. Australia publicly subsidises prescription medicines via the PBS but most medicines subsidised through the PBS are also available privately. The RPBS is a prescribed medicine subsidy scheme maintained by the Australian Department of Veterans' Affairs for their members. To maintain the quality use of medicines in Australia, the PBS and RPBS may require prescribers to phone Services Australia (the government body that manages the PBS) or lodge a request online to obtain permission before prescribing certain medicines. Such prescriptions are "authority" prescriptions [30].

5. The most recent test from the 'Pathology' table (test name and test date). This table includes the 'header row' including the test name but not the results, from pathology, radiology and other diagnostic imaging tests received.

6. A pathology test from a randomly selected, 'nominated' date from any point prior to 31st October 2020 as recorded in the 'Pathology results atom' table. This table includes the detailed (atomised) results of pathology tests when transferred in a readable format. In most cases, results for radiology and other diagnostic imaging

are not available in MedicineInsight due to the format of the results not being compatible with extraction. Additionally, results for some types of pathology tests, such as microbiology culture tests, are not available in a readable format for extraction to MedicineInsight. We assessed the agreement of test name, the returned result, and result units (e.g., 'umol/L', 'mmol', etc...).

7. Patient demographics and risk factors as recorded at 31st October 2020 in the 'Patient table' including: year of birth, sex, year of death, fact of death, and smoking status.

If no values were present in the nominated data fields, EHR reviewers recorded a missing value; EHR reviewers did not extract information from other fields such as free-text, progress notes fields.

EHR reviews

We used information obtained from the original EHRs held in the participating practices as the gold standard against which we determined agreement. EHR reviews involved a member of the research team visiting the participating practices and accessing the original EHRs at each site. EHR reviewers were provided with anonymised identifiers that were re-associated with the relevant patient names using the third-party data extraction tools installed in each practice's CIS. EHR reviews were conducted from January to March 2021. EHR reviews were undertaken by four NPS MedicineWise employees, all of whom were health professionals registered with the Australian Health Practitioner Regulation Agency and accredited for the keeping of medical records and adherence to confidentiality and privacy principles. For the prescription on a nominated date, reviewers were provided with the patient ID and the date of this prescription (but not the medicine name) and asked to record the prescription details on this date. For the test on a nominated date, reviewers were provided with the patient ID, the date of this pathology test and the test name and asked to record the test details. All reviewers were blinded to the MedicineInsight values on the variables assessed.

Post hoc EHR data cleaning

We defined agreement between original EHRs and MedicineInsight fields as exact text matches between the two data sources. We considered instances where a missing value in MedicineInsight corresponded to a missing value in the original EHR as agreement for all evaluated data fields. The 'prescribed medicine strength' and 'prescribed quantity' fields contained spelling errors and extraneous information (in both MedicineInsight and EHR data) that required review by a pharmacist/analyst prior to assessing consistency. For example, values such as '50mcg actuation' and '50 mcg actuat' [*sic*], '100 mg' and '100 mg dose', and '1.0.5 ml' and '1*0.5 ml' were considered to represent the same values. Similarly, test names, results, and result units in the collected EHR data contained spelling errors or described the same test/result/unit using slightly different terminology than that used in MedicineInsight data. A qualified General Practitioner reviewed all instances where EHR data for pathology fields did not match MedicineInsight pathology data to determine whether the discrepancy was the result of spelling errors/small differences in terminology or referred to truly different tests.

Statistical analyses

We calculated the percentage of agreement (POA), with corresponding 95% confidence intervals (95% CI) for all data fields [31]. As different fields followed different formats (e.g., categorical, continuous), we calculated additional measures of agreement and validity for specific fields. We calculated Cohen's Kappa for patient fact of death, sex, and smoking status; and for nominated prescription medicine PBS status variables. We used intraclass correlation coefficients (ICC) to assess agreement between datasets for patient year of birth and year of death [32, 33]. Finally, we calculated sensitivity, specificity, positive predictive value, and negative predictive value for recorded diagnoses (hypertension, myocardial infarction, osteoarthritis, and VTE) [34]. Where there was disagreement between data fields containing continuous values, we calculated the mean difference (MedicineInsight value – EHR value) and range of differences.

For the prescription strength, quantity, number of repeats, and PBS status data fields we only assessed agreement if the nominated medicine name for that patient matched between MedicineInsight and the EHR. Similarly, for the nominated pathology test results and test units we only assessed agreement where the test name matched between MedicineInsight and the EHR. If medical reviewers extracted values from a visit after 31st October 2020, we excluded the record for that patient, for that data field, from analysis. We also excluded any records for recent prescriptions and tests with a date before 31st October 2018. For any data fields with agreement $<$90% we further stratified the result by CIS and extraction tool. For confidentiality reasons, we cannot report results/cell counts of less than five.

We conducted all analyses using R version 3.6.2 (R Core Team, R Foundation for Statistical Computing, Vienna, Austria).

Results

Medical record reviewers extracted data for 799 patients, 796 of which we were able to link with MedicineInsight records for our analyses.

Most recent patient observations

We excluded records for 20 diastolic and 21 systolic blood pressure, six height, and seven weight observations occurring after 31st October 2020. We found excellent agreement for diastolic and systolic blood pressure observations—98% (95% CI: 97% to 99%) and 98% (95% CI: 96% to 99%), respectively (Table 1). Agreement for height and weight observations were also excellent (Table 1). The mean difference between the 14 diastolic blood pressure discrepancies was 6 mmHg (range: −20 to 35) and 4 mmHg (range: −31 to 30) for the 18 systolic blood pressure discrepancies

(no discrepancies were due to missing values). Of the 10 discrepancies for height, six were due to missing data in MedicineInsight. The mean difference for the 16 mismatched weight records was 0.5 kg (range: −5 to 13).

Historical diagnoses as recorded in the 'medical history' data field

We found excellent agreement for each of the four historical diagnoses, ranging from 97% to 99% (Table 2). All diagnoses had excellent accuracy, with sensitivities, specificities, and negative predictive values all above 90%. VTE, however, had a PPV of 57%, with MedicineInsight misclassifying nine people who did not have VTE recorded in the medical history section of the EHR, as having VTE in their MedicineInsight diagnosis field (false positives).

Most recently prescribed medicine and date of prescribing

We excluded 11 records with dates falling outside of the study period. The POAs of the remaining 785 records was 98% (95% CI: 97% to 99%) for medicine name and 97% (95% CI: 96% to 98%) for date of prescribing (Table 1). Of the 13 discrepancies for medicine name, five were due to missing values in MedicineInsight while the remaining were due to missing components of fixed-dose combinations or different medicine names. The discrepancies in date of prescribing had a mean difference of −49 days (range: −366 days to 202 days).

Prescription on nominated date—medicine name, strength, quantity, number of repeats, and PBS status

We observed POAs of 98% (95% CI: 97% to 99%), 96% (95% CI: 95% to 97%), and 99% (95% CI: 98% to 99%) for medicine name, strength and quantity, respectively (Table 1). There were 17 instances where the medicine name field in MedicineInsight did not match the EHR. Ten of 30 discrepancies for medicine strength were due to missing values in MedicineInsight, while the remaining mismatches often appeared close to matching but a single number differed (hypothetically, '100 mg/50 mg' in the EHR and '100 mg/500 mg' in MedicineInsight) which may indicate a data entry error by the medical record reviewer or a true discrepancy. POA for the PBS status of the prescription on a nominated date was 93% (95% CI: 92% to 95%), Kappa was 0.88 (95% CI: 0.85 to 0.91). Of the 52 discrepancies for this data field, 43 (83%) were due to 'Authority' prescriptions labelled as 'PBS' in MedicineInsight.

Most recent test

We excluded 14 records that occurred after 31st October 2020 or before 31st October 2018. We observed good agreement for test name and date—81% (95% CI: 78% to 83%) and 86% (95% CI: 83% to 88%), respectively (Table 1). Of the 148 discrepancies for pathology test name, less than five were due to missing values (i.e. no recorded test) in MedicineInsight while the remaining records contained differing tests in MedicineInsight and the EHR data. We found 110 discrepancies for pathology test date, six of which were due to missing values in MedicineInsight or the EHR. The mean difference for the remaining 104 records was 1 day (range: −476 to 367 days).

Pathology test on a nominated date—test name, result, and result units

We observed POAs of above 90% for the nominated test name, result and result units (Table 1). Of the 17 discrepancies for pathology test name, five were due to no test being recorded in the EHR on the randomly selected date; the remaining were due to different tests recorded. Five of the 21 discrepancies for pathology test result were due to no value in the EHR. Seven of the 28 discrepancies for result units were due to missing values in the EHR.

Patient demographics and risk factor fields

We observed high agreement for patient demographics and risk factor fields (Table 3). The largest discrepancies occurred for the patient postcode, where eight discrepancies were due to the postcode differing between the datasets and six were implausible values. Seven discrepancies for smoking status were due to missing values in MedicineInsight that were not missing in EHRs.

Discussion

Our study found excellent agreement between MedicineInsight and original EHR data fields. Patient observations, diagnoses, prescribed medicines, and demographic data fields largely matched those from original EHRs. Agreement for most recent test data fields, while not as high as other data fields, was still good. Our findings suggest that information is accurately conserved as clinical data are extracted from general practice CISs, transferred to the MedicineInsight data warehouse, transformed, cleaned, and then ultimately stored in the MedicineInsight database.

Internationally, there have been numerous validations of diagnoses data in the CPRD, but not many studies assessing other fields of CPRD data [35]. In Canada, agreement between pathology test and prescription fields in the EMRALD EHR dataset and provincial administrative records has been shown to comparable to our results for tests (ranging from 73% to 89% agreement) and prescriptions (ranging from 80% to 99% agreement) [7]. While a study of CPCSSN EHR data validated against patient medical charts found similar levels of agreement for hypertension as we observed for the medical history field in MedicineInsight data [36].

To our knowledge this is the first study of a large scale and enduring primary care data collection in Australia to assess the concordance of extracted clinical data with source data. Previous research has examined the consistency of different EHR extraction tools and found considerable differences between them [4], but not the accuracy of the data themselves. The use of data from general practice EHRs for research and policy decisions is limited but growing in Australia [11, 15]. More primary health care data collections

Table 1: Agreement between MedicineInsight and EHR data fields for patient observations, prescribed medicines, and tests

	N records agree	Percentage of agreement (95% CI)	Denominator (excluding records from outside of the study period)
Patient observations			
Diastolic blood pressure	769	98% (97% to 99%)	783
Systolic blood pressure	764	98% (96% to 99%)	782
Height	780	99% (98% to 99%)	790
Weight	773	98% (97% to 99%)	789
Most recently prescribed medicine			
Medicine name	772	98% (97% to 99%)	785
Date of prescribing	765	97% (96% to 98%)	785
Randomly selected prescribed medicine on nominated date			
Medicine name	782	98% (97% to 99%)	796
Medicine strength	755	97% (95% to 98%)	782
Medicine quantity	779	99% (99% to 99%)	782
Medicine repeats	776	99% (98% to 99%)	782
PBS status*	731	93% (92% to 95%)	782
Most recent test			
Test name	638	81% (78% to 84%)	786
(Extraction tool/CIS)			
GRHANITE™/ BP	154	79% (72% to 84%)	196
INCA™/BP	197	98% (96% to 99%)	200
GRHANITE™/MD	131	69% (62% to 75%)	191
INCA™/MD	156	78% (72% to 84%)	199
Test date	676	86% (83% to 88%)	786
(Extraction tool/CIS)			
GRHANITE™/BP	152	78% (71% to 83%)	196
INCA™/BP	197	98% (96% to 99%)	200
GRHANITE™/ MD	179	94% (89% to 96%)	191
INCA™/MD	148	74% (68% to 80%)	199
Randomly selected test on nominated date			
Test name	779	98% (97% to 99%)	796
Test result	768	99% (97% to 99%)	779
Test result units	760	98% (96% to 98%)	779

* Kappa for PBS status was: 0.88 (0.85 to 0.91). BP = Best Practice, MD = Medical Director, CIS = Clinical Information System, CI = Confidence interval.

Table 2: Agreement and accuracy of 'medical history' data field in MedicineInsight

	N records agree	Percentage of agreement (95%CI)	Sensitivity (95%CI)	Specificity (95%CI)	PPV (95%CI)	NPV) (95%CI)
Hypertension	770	97% (95% to 98%)	94% (89%–96%)	98% (96%–99%)	94% (90%–97%)	98% (96%–99%)
Myocardial infarction	794	99% (99% to 100%)	93% (66%–100%)	99% (99%–100%)	93% (66%–99%)	100% (99%–100%)
Osteoarthritis	774	97% (96% to 98%)	98% (94%–100%)	97% (95%–98%)	86% (79%–91%)	100% (99%–100%)
Venous thromboembolism	787	99% (98% to 99%)	100% (74%–100%)	99% (98%–99%)	57% (34%–78%)	100% (99%–100%)

are being developed [37] and there is an urgent need for studies validating these data, as well as studies assessing the impacts of data extraction and transformation processes.

We found that MedicineInsight extraction, transformation, and warehousing processes did not have significant impacts on the accuracy of the data. However, there were discrepancies in agreement between CIS/extraction tool combinations for tests. BP/GRHANITE and MD/INCA combinations had notably poorer agreement than MD/GRHANITE and BP/INCA combinations. The reasons for this are unclear and our sample size is too small to make generalisations based on CIS/extraction tool here. The most recent test

Table 3: Agreement between MedicineInsight and EHR data fields for patient demographics and risk factors

Patient demographic variables	N records agree	Percentage of agreement (95%CI)	Kappa (95%CI)	Intraclass correlation (95%CI)*
Patient sex	791	99% (99% to 100%)	0.99 (0.98 to 1.00)	–
Smoking status	789	99% (98% to 100%)	0.98 (0.96 to 1.00)	–
Deceased indicator (Fact of death)	791	99% (99% to 100%)	0.90 (0.80 to 1.00)	–
Year of birth	792	99% (99% to 100%)	–	0.97 (0.97 to 0.98)
Year of death	793	99% (99% to 100%)	–	0.95 (0.94 to 0.96)
Patient postcode	781	98% (98% to 100%)	–	–

*ICC models were two-way random effects models evaluating agreement with a single rater.

in MedicineInsight is based on the 'test header row' field in EHRs, which is a summary line. This field may contain entries from multiple sources (e.g. pathology labs, radiology labs, hospitals), possibly including free-text entries by GPs, and this may have caused some of the discrepancies for pathology tests. These discrepancies, and the differences between CIS/extraction tools warrant further investigation.

Other discrepancies were fewer in number and many of them may have been due to EHR fields being updated after 31st October 2020. For instance, several discrepancies for the year of death field were due to a value of '2021' entered in the EHR with no value in MedicineInsight. While the accuracy measures for diagnoses were generally excellent, and our previous research found excellent accuracy of the algorithms used to identify other conditions in MedicineInsight [28], VTE diagnoses had a poor PPV. This was due to nine false positives in the MedicineInsight diagnosis field. VTE has not previously been validated in EHR data [38] and the reason for these false positives is unclear. This issue requires further exploration.

Strengths and limitations

We used information obtained from the original EHRs held in the participating practices to assess the consistency of MedicineInsight data in this study. Information recorded in EHRs may be incomplete or inaccurate, such as in instances where healthcare providers select the wrong entry from a menu or where relevant data are manually entered with errors or in an inappropriate field [39, 40]. We did not evaluate the veracity of the EHR data, just its agreement with MedicineInsight data values. We considered instances where values were missing in both MedicineInsight and EHRs as agreement. Our EHR data were extracted by hand, by trained medical record reviewers. It is possible that some data were not correctly extracted by the medical record reviewers. We attempted to remedy this situation where possible (e.g. spelling errors) and have highlighted where we believe incorrect information may have been entered. For specific variables, EHR reviewers were instructed to record data from the most recent visit between 31st October 2018 and 31st October 2020, however, several records were extracted outside of these dates and, therefore, could not be matched with MedicineInsight values. We excluded these values from analyses as they would have biased our agreement results. Resource constraints limited our study to four general practice sites - a small sample that may not be entirely representative of other general practices. Further, to be eligible for inclusion in our study, practices had to meet criteria related to patient load, location (metropolitan Sydney and Melbourne) and interest in engaging with the MedicineInsight program. These criteria may have led to the inclusion of practices that are not representative of all MedicineInsight practices in terms of data extraction and transmission. We selected practices representing the different combinations of available CISs and data extraction software tools to mitigate these impacts. Different practices use different external vendors (e.g. pathology laboratories) who may return improperly coded results to the practice CIS, resulting in practice-level differences that may also influence reliability [29].

Conclusions

Overall, our study shows good agreement between the majority of MedicineInsight data and those from original EHRs, suggesting MedicineInsight data extraction and warehousing procedures accurately conserved the clinical data in our sample. However, there was poor PPV for VTE diagnosis and this warrants further investigation. There were also small amounts of missing data and potentially erroneous entries - whether due to inaccurate data entry by the healthcare provider or extraction errors. Given the other agreement results our study found, we would expect the impacts of any discrepancies to be minimal, but researchers should be aware of these issues when using MedicineInsight. The difference in results observed between CISs and data extraction tools warrants additional exploration.

Acknowledgements

We thank the staff and patients of the four general practices participating in this study. We are also grateful to Melissa Chapman, Jing Ye, Jason Mak, Sindu Murugathas, Reya Bokshi, Lisa Quick, Jill Thistlethwaite and Rob Mina for their individual contributions to this research.

Funding

This study was funded by the Australian Government Department of Health. The funding body had no role in the

design of the study, data collection, analysis or interpretation, nor in writing the manuscript. BD and AH are supported by the National Health and Medical Research Council (NHMRC) Centre of Research Excellence in Medicines Intelligence (ID: 1196900).

Statement on conflicts of interest

All authors are employees or collaborators of NPS MedicineWise, the custodian of the MedicineInsight data. The Centre for Big Data Research in Health, UNSW Sydney received funding from AbbVie Australia in 2020 to conduct research, unrelated to the present study.

Ethics statement

NPS MedicineWise was granted ethics approval for the standard operations and uses of the MedicineInsight database in December 2017. This program approval was given by the Royal Australian College of General Practitioners (RACGP) National Research and Evaluation Ethics Committee (NREEC 17-017). Additional ethics approval for our study was granted by the RACGP NREEC (NREEC 19-010) on 28 October 2020, with this approval ratified by the University of New South Wales Human Research Ethics Committee (19-010/AH03552). Our study also received approval from the MedicineInsight Data Governance Committee on the 12th August 2020 (2020–022).

References

1. Campanella P, Lovato E, Marone C, Fallacara L, Mancuso A, Ricciardi W, et al. The impact of electronic health records on healthcare quality: a systematic review and meta-analysis. Eur J Public Health. 2016;26(1):60–4. https://doi.org/10.1093/eurpub/ckv122

2. Nguyen L, Bellucci E, Nguyen LT. Electronic health records implementation: an evaluation of information system impact and contingency factors. Int J Med Inform. 2014;83(11):779–96. https://doi.org/10.1016/j.ijmedinf.2014.06.011

3. Gentil ML, Cuggia M, Fiquet L, Hagenbourger C, Le Berre T, Banâtre A, et al. Factors influencing the development of primary care data collection projects from electronic health records: a systematic review of the literature. BMC Med Inform Decis Mak. 2017;17(1):139. https://doi.org/10.1186/s12911-017-0538-x

4. Liaw ST, Taggart J, Yu H, de Lusignan S. Data extraction from electronic health records - existing tools may be unreliable and potentially unsafe. Aust Fam Physician. 2013;42(11):820–3.

5. Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, van Staa T, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). International Journal of Epidemiology. 2015;44(3):827–36. https://doi.org/10.1093/ije/dyv098

6. The Health Improvement Network. What is THIN(R) data? 2021 [Available from: https://www.the-health-improvement-network.com/en/#what-is-thin.

7. Tu K, Mitiku TF, Ivers NM, Guo H, Lu H, Jaakkimainen L, et al. Evaluation of Electronic Medical Record Administrative data Linked Database (EMRALD). Am J Manag Care. 2014;20(1):e15–21.

8. Birtwhistle RV. Canadian Primary Care Sentinel Surveillance Network: a developing resource for family medicine and public health. Can Fam Physician. 2011;57(10):1219–20.

9. United States Department of Veterans Affairs. Corporate Data Warehouse 2021 [Available from: https://www.hsrd.research.va.gov/for_researchers/vinci/cdw.cfm.

10. McInnes DK, Saltman DC, Kidd MR. General practitioners' use of computers for prescribing and electronic health records: results from a national survey. Med J Aust. 2006;185(2):88–91. https://doi.org/10.5694/j.1326-5377.2006.tb00479.x

11. Canaway R, Boyle DI, Manski-Nankervis JE, Bell J, Hocking JS, Clarke K, et al. Gathering data for decisions: best practice use of primary care electronic records for research. Med J Aust. 2019;210 Suppl 6(Suppl Suppl 6):S12–s6. https://doi.org/10.5694/mja2.50026

12. Youens D, Moorin R, Harrison A, Varhol R, Robinson S, Brooks C, et al. Using general practice clinical information system data for research: the case in Australia. Int J Popul Data Sci. 2020;5(1):1099-. https://doi.org/10.23889/ijpds.v5i1.1099

13. 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. Int J Epidemiol. 2019;48(6):1741-h. https://doi.org/10.1093/ije/dyz147

14. Health Communication Network Limited. Practice Incentives Program (PIP) eHealth Incentive: Requirement 3—Data Records and Clinical Coding. 2014.

15. NPS MedicineWise. Approved Projects Using MedicineInsight 2021 [Available from: https://www.nps.org.au/medicine-insight/approved-projects-using-medicineinsight-data.

16. Busingye D, Daniels B, Brett J, Pollack A, Belcher J, Chidwick K, et al. Patterns of real-world opioid prescribing in Australian general practice (2013–2018). Australian Journal of Primary Health. 2021;27:416–24. https://doi.org/10.1071/PY20270

17. Black-Tiong S, Gonzalez-Chica D, Stocks N. Trends in long-term opioid prescriptions for musculoskeletal conditions in Australian general practice: a national longitudinal study using MedicineInsight, 2012-2018. BMJ open. 2021;11(4):e045418. https://doi.org/10.1136/bmjopen-2020-045418

18. Totterdell J, Phillips A, Glover C, Chidwick K, Marsh J, Snelling T, et al. Safety of live attenuated herpes zoster vaccine in adults 70–79 years: A self-controlled case series analysis using primary care data from Australia's MedicineInsight program. Vaccine. 2020;38(23):3968–79. https://doi.org/10.1016/j.vaccine.2020.03.054

19. De Oliveira Bernardo C, González-Chica DA, Chilver M, Stocks N. Influenza immunisation coverage from 2015 to 2017: A national study of adult patients from Australian general practice. Vaccine. 2019;37(31):4268–74. https://doi.org/10.1016/j.vaccine.2019.06.057

20. de Oliveira Costa J, Gianacas C, Beard F, Gonzalez-Chica D, Chidwick K, Osman R, et al. Cumulative annual coverage of meningococcal B vaccination in Australian general practice for three at-risk groups, 2014 to 2019. Human Vaccines & Immunotherapeutics. 2021:1–10. https://doi.org/10.1080/21645515.2021.1923349

21. Chidwick K, Kiss D, Gray R, Yoo J, Aufgang M, Zekry A. Insights into the management of chronic hepatitis C in primary care using MedicineInsight. Australian journal of general practice. 2018;47(9):639–45. https://doi.org/10.31128/ajgp-02-18-4482

22. Businge D, Chidwick K, Simpson V, Dartnell J, G JD, Balcomb A, et al. The changing characteristics of patients with chronic hepatitis C prescribed direct acting antiviral medicines in general practice since listing of the medicines on the Australian Pharmaceutical Benefits Scheme. JGH Open. 2021;5(7):813–9. https://doi.org/10.1002/jgh3.12593

23. Healthdirect Australia. National Health Services Directory Sydney: Healthdirect Australia; 2019 [Available from: https://studio.healthmap.com.au/.

24. Deeble Institute. Deeble Institute Issues Brief No. 18: Reality check - reliable national data from general practice electronic health records 2016 [Available from: https://ahha.asn.au/publication/issue-briefs/deeble-institute-issues-brief-no-18-reality-check-reliable-national-data.

25. University of Melbourne Health Informatics Centre. What is GRHANITE? 2021 [Available from: https://www.grhanite.com/technologies/.

26. Precedence Health Care. How INCA Works 2021 [Available from: https://precedencehealthcare.com/inca/.

27. World Health Organisation. ATC: Structure and principles 2021 [Available from: https://www.whocc.no/atc/structure_and_principles/.

28. Havard A, Manski-Nankervis J-A, Thistlethwaite J, Daniels B, Myton R, Tu K, et al. Validity of algorithms for identifying five chronic conditions in MedicineInsight, an Australian national general practice database. BMC Health Services Research. 2021;21(1):551. https://doi.org/10.1186/s12913-021-06593-z

29. Peiris D, Agaliotis M, Patel B, Patel A. Validation of a general practice audit and data extraction tool. Aust Fam Physician. 2013;42(11):816–9.

30. Mellish L, Karanges EA, Litchfield MJ, Schaffer AL, Blanch B, Daniels BJ, et al. The Australian Pharmaceutical Benefits Scheme data collection: a practical guide for researchers. BMC Res Notes. 2015;8:634. https://doi.org/10.1186/s13104-015-1616-8

31. Jakobsson U, Westergren A. Statistical methods for assessing agreement for ordinal data. Scandinavian Journal of Caring Sciences. 2005;19(4):427–31. https://doi.org/10.1111/j.1471-6712.2005.00368.x

32. Ranganathan P, Pramesh CS, Aggarwal R. Common pitfalls in statistical analysis: Measures of agreement. Perspect Clin Res. 2017;8(4):187–91. https://doi.org/10.4103/picr.PICR_123_17

33. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med. 2016;15(2):155–63. https://doi.org/10.1016/j.jcm.2016.02.012

34. Akobeng AK. Understanding diagnostic tests 1: sensitivity, specificity and predictive values. Acta Paediatrica. 2007;96(3):338–41. https://doi.org/10.1111/j.1651-2227.2006.00180.x

35. Herrett E, Thomas SL, Schoonen WM, Smeeth L, Hall AJ. Validation and validity of diagnoses in the General Practice Research Database: a systematic review. British Journal of Clinical Pharmacology. 2010;69(1):4–14. https://doi.org/10.1111/j.1365-2125.2009.03537.x

36. 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. Ann Fam Med. 2014;12(4):367–72. https://doi.org/10.1370/afm.1644

37. Australian Institute of Health and Welfare. Developing a National Primary Health Care Data Asset: consultation report. Canberra: AIHW; 2019.

38. McBrien KA, Souri S, Symonds NE, Rouhi A, Lethebe BC, Williamson TS, et al. Identification of validated case definitions for medical conditions used in primary care electronic medical record databases: a systematic review. Journal of the American Medical Informatics Association. 2018;25(11):1567–78. https://doi.org/10.1370/afm.1644

39. NPS MedicineWise. Validation of the MedicineInsight database: completeness, generalisability and plausibility. https://www.nps.org.au/medicine-insight/using-medicineinsight-data; 2020.

40. Madden JM, Lakoma MD, Rusinak D, Lu CY, Soumerai SB. Missing clinical and behavioral health data in a large electronic health record (EHR) system. J Am Med Inform Assoc. 2016;23(6):1143–9. https://doi.org/10.1093/jamia/ocw021

Supplementary Appendix 1. Terms used by medical record reviewers to identify diagnoses.

The patient was considered to have the indicated condition if one of the terms (any spelling or phrasing variant) listed below was present in the EHR:

Hypertension

ANTIHYPERTENSIVE AGENT PRESCRIPTION
BLOOD PRESSURE LABILE
BLOOD PRESSURE REVIEW
BP LABILE
BP UNSTABLE
DIASTOLIC HYPERTENSION
ESSENTIAL HYPERTENSION
HBP
HIGH BLOOD PRESSURE
HIGH BP
HT (HYPERTENSION)
H/T
HYPERTENSION
HYPERTENSION - CONTROLLED
HYPERTENSION - ISOLATED SYSTOLIC
HYPERTENSION - LABILE
HYPERTENSION - LIFE STYLE MANAGEMENT
HYPERTENSION - MALIGNANT
HYPERTENSION - PREGNANCY
HYPERTENSION - RENOVASCULAR
HYPERTENSION - UNSTABLE
HYPERTENSION IN PREGNANCY
HYPERTENSION REVIEW
HYPERTENSION, ISOLATED SYSTOLIC
HYPERTENSION, DIASTOLIC
HYPERTENSION, ESSENTIAL
HYPERTENSION, MALIGNANT
HYPERTENSION, RENOVASCULAR
HYPER TENSION
ISOLATED SYSTOLIC HYPERTENSION
LABILE BLOOD PRESSURE
LABILE BP
LABILE HYPERTENSION
MALIGNANT HYPERTENSION
PIH
PREGNANCY INDUCED HYPERTENSION
PRIMARY HYPERTENSION
RENAL HYPERTENSION
RENOVASCULAR HYPERTENSION
REVIEW – BP
SEVERE REFRACTORY HYPERTENSION

Myocardial infarction

ACUTE MYOCARDIAL INFARCTION
AMI
ANTERIOR MYOCARDIAL INFARCT
ANTEROLATERAL MYOCARDIAL INFARCT
DRESSLER'S SYNDROME
HEART ACK +YEAR
HEART ATTACK
HISTORY OF HEART ATTACK
HISTORY OF INFARCT
HISTORY OF MI
INFARCT +YEAR
INFERIOR MYCARDIAL INFARCTION
INFERIOR MYCARDIAL INFARCTION
MI
MI +YEAR
MYOCARDIAL INFARCT - SILENT
MYOCARDIAL INFARCTION
MYOCARDIAL INFARCTION - ANTEROLATERAL
MYOCARDIAL INFARCTION - INFERIOR
MYOCARDIAL INFARCTION - POSTERIOR
MYOCARDIAL INFARCTION - SUBENDOCARDIAL
MYOCARDIAL INFARCTION - SUPERIOR
MYOCARDIAL INFARCTION, ANTERIOR
MYOCARDIAL INFARCTION, ANTEROLATERAL
MYOCARDIAL INFARCTION, INFERIOR
MYOCARDIAL INFARCTION, NON STEMI
MYOCARDIAL INFARCTION, POSTERIOR
MYOCARDIAL INFARCTION, STEMI
MYOCARDIAL INFARCTION, SUBENDOCARDIAL
MYOCARDIAL INFARCTION, SUPERIOR
NON ST ELEVATION MYOCARDIAL INFARCTION
NSTEMI
OLD HEART ATTACK
OLD INFARCT
OLD MI
PAST HEART ATTACK
PAST INFARCT
PAST MI
POSTERIOR MYOCARDIAL INFARCT
POSTERIOR MYOCARDIAL INFARCTION
POSTMYOCARDIAL INFARCTION SYNDROME
POSTPERICARDIOTOMY SYNDROME
SILENT MYOCARDIAL INFARCTION
ST ELEVATION MYOCARDIAL INFARCTION
ST ELEVATON MYOCARDIAL INFARCTION
SUBENDOCARDIAL INFARCT
SUBENDOCARDIAL MYOCARDIAL INFARCT
SUBENDOCARDIAL MYOCARDIAL INFARCT
SUPERIOR MYOCARDIAL INFARCT
SUPERIOR MYOCARDIAL INFARCTION

Osteoarthritis

ANEURYSM-OSTEOARTHRITIS SYNDROME
ANKLE OSTEOARTHRITIS
ANKYLOSING SPONDYLITIS
ARTHRITIS - OSTEO
CERVICAL - OSTEO ARTHRITIS
CERVICAL SPINE OSTEOARTHRITIS
ELBOW OSTEOARTHRITIS
GENERALISED OSTEOARTHRITIS
HALLUX RIGIDUS
HIP OSTEOARTHRITIS
HIP OSTEOARTHROSIS
KNEE OSTEOARTHRITIS
KNEE OSTEOARTHROSIS
LUMBAR - OSTEO ARTHRITIS

LUMBAR SPINE OSTEOARTHRITIS
MIDFOOT OSTEOARTHRITIS
OA
OA (OSTEOARTHRITIS)
OSTEOARTHRITIS
OSTEOARTHRITIS - ANKLE
OSTEOARTHRITIS - ELBOW
OSTEOARTHRITIS - FINGERS
OSTEOARTHRITIS - GLENOHUMERAL JOINT
OSTEOARTHRITIS - HANDS
OSTEOARTHRITIS - HIP
OSTEOARTHRITIS - KNEE
OSTEOARTHRITIS - NECK
OSTEOARTHRITIS - SHOULDER
OSTEOARTHRITIS - SPINE
OSTEOARTHRITIS OF 1ST CARPOMETACARPAL JOINT
OSTEOARTHRITIS OF 1ST CARPO-METACARPAL JOINT
OSTEOARTHRITIS OF 1ST METATARSOPHALANGEAL JOINT
OSTEOARTHRITIS OF ANKLE
OSTEOARTHRITIS OF CERVICAL SPINE
OSTEOARTHRITIS OF ELBOW
OSTEOARTHRITIS OF FINGERS
OSTEOARTHRITIS OF FOOT
OSTEOARTHRITIS OF HAND
OSTEOARTHRITIS OF HIP
OSTEOARTHRITIS OF KNEE
OSTEOARTHRITIS OF LUMBAR SPINE
OSTEOARTHRITIS OF NECK
OSTEOARTHRITIS OF SACROILIAC JOINTS
OSTEOARTHRITIS OF SHOULDER
OSTEOARTHRITIS OF THE PATELLOFEMORAL JOINT
OSTEOARTHRITIS OF THE STERNOCLAVICULAR JOINT
OSTEOARTHRITIS OF THORACIC SPINE
OSTEOARTHRITIS OF TMJ
OSTEOARTHRITIS OF WRIST
OSTEOARTHRITIS, GENERALISED
OSTEOARTHROSIS
PATELLOFEMORAL OSTEOARTHRITIS
SACROILIAC JOINT ARTHRITIS
SHOULDER OSTEOARTHRITIS
SPONDYLOSIS
STERNO-CLAVICULAR OSTEOARTHRITIS
THORACIC - OSTEO ARTHRITIS
WEAR AND TEAR ARTHRITIS
WRIST OSTEOARTHRITIS

Venous thromboembolism

DEEP VENOUS THROMBOSIS
DVT
D.V.T
THROMBOSIS - DEEP VEIN
VTE
EMBOLISM - PULMONARY
EMBOLISM, PULMONARY
PE
PULMONARY EMBOLISM
PULM EMB
PULMONARY EMB
PULMONARY EMBOLISM - SADDLE TYPE
SADDLE PULMONARY EMBOLISM

Note:
The abbreviation 'PE' can refer to many terms such as 'Physical Examination' , 'Pre-eclampsia', 'Premature ejaculation' etc. 'PE' was only assumed to refer to 'Pulmonary embolism' when reported with a related condition or other qualifier, or in conjunction with an anticoagulant e.g.:
• Flagged as 'Pulmonary embolism' – WARFARIN FOR PE', 'CLEXANE FOR PE', 'DVT AND PE', 'SADDLE PE', 'LEFT LUNG PE', 'UNPROVOKED PE', 'MALIGNANT PE'
• Not flagged as 'Pulmonary embolism' – 'PE', 'POST OP PE', 'PE UNREMARKABLE', 'HYERTENSION POSSIBLE PE', 'PE ED LUTS'