MEDICINEINSIGHT

Validation of the MedicineInsight general practice database: the accuracy of death recording in the MedicineInsight data compared with the National Death Index in Australia.

Australian Government Department of Health

December 2021 version 1.0

Independent, not-for-profit and evidence-based, NPS MedicineWise enables better decisions about medicines, medical tests and other health technologies.

NPS MEDICINEWISE

Level 7/418A Elizabeth St
Surry Hills NSW 2010
PO box 1147
Strawberry Hills NSW 2012

P. 02 8217 8700
F. 02 9211 7578
info@nps.org.au
www.nps.org.au

Suggested citation

Myton R, Pollack A, Havard A, Belcher J, Annear K, Chidwick K. MedicineInsight report: Validation of the MedicineInsight database: the accuracy of death recording in the MedicineInsight general practice data compared with the National Death Index in Australia. Sydney: NPS MedicineWise, 2021.

Acknowledgments

This project is funded by the Australian Government Department of Health. NPS MedicineWise is responsible for the design, analysis and publication of the results of the project. The Australian Government Department of Health was not involved in the analysis and interpretation of the data.

We are grateful to the general practices and general practitioners who participate in MedicineInsight and the patients whose de-identified data makes this work possible. We would also like to acknowledge NPS MedicineWise staff, particularly Lisa Quick, Doreen Busingye, Yuen Ai Lee and Jill Thistlethwaite, who contributed to this report.

Contents

Executive summary 4

Key findings 5

Discussion 6

Recommendations 6

1. Background 8

1.1. MedicineInsight program 8

1.2. Assessment of the validity of MedicineInsight data 8

1.3. Focus of the current study – death recording in MedicineInsight 9

1.4. The NDI 10

1.5. Ethics and data governance approvals for the use of linked MedicineInsight and NDI data in this study 10

2. Aims and methods 11

2.1. Aim 11

2.2. Study design 11

2.3. Data linkage process 11

2.4. Study period 14

2.5. Study cohort 15

2.6. Definitions 16

2.7. Data analysis and reporting 20

3. Results 22

Key findings 22

3.1. Study cohorts 22

3.2. Fact of death 24

3.3. Date of death 31

3.4. Representativeness of the linked MedicineInsight–NDI study population 34

3.5. Exploratory analysis to identify duplicate patients 37

Guide to interpreting the data 40

References 41

Appendix 1 42

EXECUTIVE SUMMARY

MedicineInsight is a database held by NPS MedicineWise containing de-identified electronic health records (EHRs) from over 700 Australian general practices. MedicineInsight data are used for quality improvement activities and for research and evaluation, program design and policy development. The extent to which the findings of analyses of MedicineInsight data are a true reflection of general practice activities and patient health, and are trusted by clinicians, policymakers and researchers, depends on the quality and completeness of the included data.

This study assessed whether recording of deaths in the MedicineInsight dataset is consistent with information obtained through individual privacy preserving record linkage (PPRL) with the National Death Index (NDI) – the 'gold standard' reference source for deaths in Australia. The NDI is a database held by the Australian Institute of Health and Welfare (AIHW) containing death registration data (fact, date and coded cause of death based on the death certificate) for all deaths that have occurred in Australia since 1980.

Information on death is important in descriptive epidemiology for defining the end of a patient's 'follow-up time' (ie, time present in a longitudinal dataset or open cohort such as MedicineInsight), especially for analyses of mortality and for studies assessing events at the end of life. Death and survival are important outcomes in analytic epidemiology, particularly for post-market surveillance of therapeutics, including effectiveness and safety studies. It is not clear whether all deaths are recorded in MedicineInsight or whether date of death can be accurately estimated from the information available. This is because general practitioners may not receive information about their deceased patients if they did not complete the death certificate; there may be delays in notification and there will be differences in recording of deaths between practices and clinical information systems.

To establish whether the death information available in MedicineInsight data can be validly used without reference to an external data source, this study addressed three aims:

1. To examine the validity of the MedicineInsight algorithm for fact of death.
2. To examine the validity of the MedicineInsight algorithm for date of death.
3. To assess potential variation in validity of the above algorithms over time, and among regular and infrequent attenders.

A PPRL between MedicineInsight and NDI data using 'Bloom filters' was undertaken by the Curtin Data Linkage (CDL) unit at Curtin University. Bloom filters enable privacy preserving linkage by encoding patient identifiers 'at source' into a non-identifiable format that can be extracted and linked probabilistically to identifiers from other datasets, which have been encoded using exactly the same process. Currently only MedicineInsight practice sites utilising the INCA extraction tool (not the GRHANITE extraction tool) can generate the Bloom filters necessary for linkage.

Key findings

The linked MedicineInsight–NDI study cohorts

- This analysis was conducted separately for five MedicineInsight–NDI linked patient cohorts in consecutive 2-year time periods: 2011–12; 2013–14; 2015–16; 2017–18; and 2019–20. Results were also combined for 2011–2020.
- Eligible general practice sites were those included in both the linked MedicineInsight–NDI dataset* (239 general practice sites) and the August 2021 MedicineInsight data download (195 general practice sites with a complete data extract and that met data quality criteria, from 239 sites).
- Eligible patients were those with at least one clinical encounter during one or more of the five consecutive 2-year study periods at an eligible general practice site. 'Regular attenders' had three or more clinical encounters during a 2-year study period and 'infrequent attenders' had 1–2 clinical encounters during a 2-year study period.
- The number of eligible general practices ranged from 156 in 2011–12 to 195 in 2019–20.
- The number of patients eligible for the five consecutive 2-year study periods ranged from 821,707 (444,696 regular attenders and 377,011 infrequent attenders) in 2011–12 to approximately 1.36 million (789,629 regular attenders and 568,006 infrequent attenders) in 2019–20. The total (2011–2020) linked population included 3.07 million patients (1.69 million regular attenders and 1.38 million infrequent attenders).

Fact of death

- The percentage of agreement (PoA) between MedicineInsight deaths and those in the NDI was excellent across all years and all patients (regular and infrequent attenders) – all PoA were above 99.0%.
- Accuracy for fact of death was mixed, with excellent specificity, positive predictive value (PPV) and negative predictive value (NPV) but poor sensitivity.
- Accuracy for fact of death was better among regular than infrequent attenders and didn't change substantially over time (2011 to 2020).
- For regular attenders (2011 to 2020), the agreement on fact of death was excellent (PoA 99% (95% confidence interval [CI]: 99% to 100%) and accuracy was mixed: sensitivity 66% (95% CI: 62% to 70%); specificity 100% (95% CI: 100% to 100%); PPV 96% (95% CI: 96% to 97%); and NPV 99% (95% CI: 99% to 99%).
    - Between 2011 and 2020, 62,031 regular attenders had a 'gold standard' record of death (3.7% of 1.69 million patients) in the NDI data; 40,930 (66.0%) of these deaths were also recorded in MedicineInsight (true positive) and 21,101 (34.0%) were not (false negative).
    - Between 2011 and 2020, 42,549 regular attenders had a death recorded (2.5% of 1.69 million patients) in MedicineInsight; 40,930 (96.2%) of these deaths were also recorded in NDI data (true positive) and 1619 (3.8%) were not (false positive).

* Only MedicineInsight participating practices utilising the INCA extraction tool were included as currently only the INCA extraction tool accommodates the 'Bloom filters' which are required for linkage.

Date of death

- The MedicineInsight inferred date of death was in agreement, within ±30 days, for 74.4% of 43,747 patients with a record of death in both MedicineInsight and NDI data during the 10-year study period (2011 to 2020).
- The accuracy of the MedicineInsight inferred death date algorithm (±30 days) increased moderately over time from 71.6% for the 2011–12 cohort to 77.1% in the 2019–20 cohort.
- Agreement on death date (±30 days) was higher among regular attenders (75.4%) and lower among infrequent attenders (60.4%) over the 10-year study.

Discussion

The deaths that were recorded in MedicineInsight could be validated against NDI deaths, with a relatively small number of 'false positive' death records resulting in an excellent PPV (95.8% overall and 96.2% for regular attenders). However, it is clear there is underreporting of deaths in MedicineInsight compared with NDI data, with the MedicineInsight deceased algorithm returning a high number of 'false negative' death records, resulting in a poor sensitivity (59.5% overall and 66.0% for regular attenders).

Despite the poor sensitivity of the MedicineInsight death algorithm the PoA was high because few patients overall (2.4% of all patients and 3.7% of regular attenders; 2011–2020) died according to NDI data, meaning the large majority of MedicineInsight patients were concordant for death recording.

Recommendations

- For studies where death is an important outcome, MedicineInsight cannot be reliably used without linkage to NDI data or other external data sources such as the state and territory death registers. Examples of these types of studies include estimating mortality rates or survival among population groups and assessing the association between exposure to a therapeutic product and death.
- Both the high PPV and specificity of deaths recorded in MedicineInsight, and the relatively high concordance on date of death (±30 days) between MedicineInsight and NDI records indicate MedicineInsight could be used for end-of-life studies, which describe the management of patients in the years prior to their death, provided these patients are representative of all deceased patients.
- The PoA was excellent, because only a small proportion of patients die during the usual reporting period (often 1 to 2 years) of MedicineInsight studies. This provides reassurance that for most descriptive epidemiological studies, such as studies on the prevalence and incidence of common chronic conditions and the use of therapeutics, MedicineInsight data can be confidently used without reference to NDI data. However, more caution may be required in studies involving aged populations and high-risk conditions (eg, heart failure, severe chronic kidney disease). An examination of the validity of MedicineInsight algorithms for death among older patients (eg, 70+ years) would help determine the importance of linkage in these situations.
- This study only examined the validity of death identification during periods of attendance at MedicineInsight practices, and results should not be generalised to deaths occurring long after a patient's last-recorded encounter. Periodic (eg, 6 monthly) regular linkage between

MedicineInsight and the NDI data is recommended to enable studies to be readily conducted, without delays due to data access, where fact of death is an important outcome. To use the linked datasets, individual approvals from both AIHW and NPS MedicineWise would be required. A shared/streamlined approval process for such projects could be explored with AIHW.

- Currently, only some MedicineInsight practices (those using the INCA extraction tool) can be linked using our preferred PPRL methodology. Expanding the number of MedicineInsight practices eligible for linkage should be a priority to improve the sample size and resulting power of future studies requiring linked data. The MedicineInsight–NDI linked dataset included 239 MedicineInsight practice sites (195 were also in the August MedicineInsight download used in this study), which is under half the usual number of quality practice sites included in MedicineInsight reports.
- To comprehensively assess the quality of the PPRL, further validation studies are recommended. For example, duplicate patients identified through linkage could be validated as true duplicates 'at source' by reidentifying patients back at the practice and checking the EHR.

1. BACKGROUND

1.1. MedicineInsight program

MedicineInsight is a large-scale database containing de-identified electronic health records (EHRs) from almost 700 participating general practices across Australia. MedicineInsight was initially established by NPS MedicineWise in 2011, with core funding from the Australian Government Department of Health. It collects general practice data to support quality improvement in Australian primary care and post-market surveillance of medicines.

MedicineInsight uses third-party data extraction tools (GeneRic Health Network Information Technology for the Enterprise [GRHANITE],[1] and Precedence Health Care's INCA[2]) which de-identify, extract and securely transmit whole-of-practice data from within each practice's clinical information system (CIS); either Best Practice or Medical Director. A whole-of-practice data collection, containing all available historic and current EHRs, is conducted when a practice joins MedicineInsight. Fields potentially containing identifying information, such as progress notes and correspondence, are not included in the extract. The extraction tool collects incremental data regularly, resulting in an updated longitudinal database in which patients attending each practice can be tracked over time. Currently only the INCA extraction tool (not GRHANITE) can generate the "Bloom filters" (see Appendix 1) necessary for linkage.

Patient identifying data such as name, date of birth and address are not extracted, although year of birth and postcode are, enabling the calculation of age, geographical location, remoteness and Socio-Economic Indexes for Areas. Extracted data include patient demographics (year of birth, sex, postcode) and clinical data entered directly by healthcare professionals (diagnoses, observations, tests performed, medicines prescribed). Each patient is assigned a unique number that allows all the records held in the database to be linked to the associated patient.

MedicineInsight data are only used and shared consistent with the principles of public good, including contributing to improving health outcomes for Australians.

Further information is available online: https://www.nps.org.au/medicine-insight

1.2. Assessment of the validity of MedicineInsight data

The extent to which the findings from MedicineInsight data are a true reflection of general practice activities and patient health, and are trusted by clinicians, policymakers and researchers, depends on the quality and completeness of the data. MedicineInsight reflects everyday health care provided to patients within a sample of practices across Australia. MedicineInsight data are real-world data entered into the CIS by practice staff for the purposes of providing clinical care and administrative activities within the practice, and not for the purpose of research.

NPS MedicineWise works with practices to improve data quality in multiple ways. For example, when practice quality improvement reports are developed as part of the implementation of national therapeutic educational programs, the quality of the data is checked with sentinel practices to ensure there is correct identification of patients, medicines, tests, conditions and other relevant data elements.

Previous research examining the validity of MedicineInsight algorithms (flags) for identifying five medical conditions (anxiety, asthma, depression, osteoporosis and type 2 diabetes) found these measures were highly accurate when compared with gold-standard EHRs.[2] Comprehensive information on the completeness, generalisability and plausibility of the MedicineInsight data[3] is available online: https://www.nps.org.au/assets/MedicineInsight-Validation-completeness-representativeness-plausibility_2020.pdf

1.3. Focus of the current study – death recording in MedicineInsight

To establish whether the death information available in MedicineInsight data can be validly used without reference to an external data source, this study addressed three aims:

1. To examine the validity of the MedicineInsight algorithm for fact of death.
2. To examine the validity of the MedicineInsight algorithm for date of death.
3. To assess potential variation in validity for the above algorithms over time, and among regular and infrequent attenders.

Death and survival are important outcomes in post-market surveillance, effectiveness and safety studies.[4,5] Information on the date of death is also important for defining patient follow-up time in cohort studies (ie, the end of a patient's time present in a longitudinal dataset) in observational epidemiology, especially for analyses of mortality and for studies assessing events at the end of life.[6]

It is not clear whether all deaths are recorded in MedicineInsight or whether date of death can be accurately estimated from the information available. This study examines the accuracy of fact of death and date of death identification in MedicineInsight through individual level linkage with National Death Index (NDI) data. The data recorded in NDI are considered the gold standard. This external validation will assess whether linkage to external data sources is required when death is an important outcome of MedicineInsight post-market surveillance and safety studies.

General practitioners (GPs) may not routinely receive information about their deceased patients if they did not complete the death certificate. In Australia it is the responsibility of the GP, the treating doctor in hospital or the Coroner's office to complete the death certificate including the cause of death. The medical practitioner responsible for the deceased person's medical care during their last illness or immediately before death, or who examined the body of the deceased person after death, can complete the death certificate. The practitioner must be 'comfortably satisfied' as to the cause of death, with no other circumstances present that require the death to be reported to the Coroner.[7,8]

Identifying deceased patients is further complicated because general practice CIS allow the recording of death information via multiple methods. Changes to the software over time or when converting between systems may affect the completeness of records.

NPS MedicineWise has developed an algorithm to identify multiple potential records of death from the patient's health records and combine this information to attempt to identify the best estimate for the date of death. This has not been externally validated (see Section 2.6 for details of the algorithm).

1.4. The NDI

The NDI is a database developed and maintained by the Australian Institute of Health and Welfare (AIHW). The database lists deaths that have occurred in Australia since 1980. It is an invaluable tool for epidemiologists and clinicians in following up research cohorts using record linkage. The NDI death registration data contain the date and coded cause of death for the population of Australia based on the death certificate, and are considered the gold standard. Since 1997 all causes of death, including the 'underlying cause of death' and 'other causes of death', are classified using the International Classification of Diseases (ICD)-10. Fact-of-death data are usually made available to AIHW within 2 months and cause of death data may take about 18 months to become available to AIHW. NDI cause of death information was not required, and therefore not requested, for this project.

1.5. Ethics and data governance approvals for the use of linked MedicineInsight and NDI data in this study

In December 2017, NPS MedicineWise was granted ethics approval for the standard operations and uses of the MedicineInsight database by NPS MedicineWise. 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 this specific project was granted by the AIHW Ethics Committee on 28 October 2020 (EO2020/4/1198). The project also received approval from the MedicineInsight Data Governance Committee on 12 August 2020 (2020-023).

2. AIMS AND METHODS

2.1. Aim

To establish whether the death information available in MedicineInsight data can be validly used without reference to an external data source, this study addressed three aims:

1. To examine the validity of the MedicineInsight algorithm for fact of death.
2. To examine the validity of the MedicineInsight algorithm for date of death.
3. To assess potential variation in validity for the above algorithms over time, and among regular and infrequent attenders.

2.2. Study design

This was a validation study using linked data to compare the identification of deaths in MedicineInsight to the Australian NDI, based on the August 2021 MedicineInsight data download.

2.3. Data linkage process

A privacy preserving record linkage (PPRL) using "Bloom filters" (see Appendix 1) was undertaken by the Curtin Data Linkage (CDL) unit at Curtin University. Bloom filters enable PPRL by encoding patient identifiers 'at source' (into a sequence of 1s and 0s called the Bloom filter hash). These identifiers can be extracted and linked probabilistically to identifiers from other datasets, which have been encoded using exactly the same process. Currently only MedicineInsight practice sites using the INCA extraction tool can generate the Bloom filters necessary for linkage. The patient-level identifiers used to generate the Bloom filters 'at source' from NDI data and at the included MedicineInsight practice sites included: First name, Middle name, Surname, Date of Birth (DOB) Year, DOB Month, DOB Day, Sex, Address, Suburb, Postcode, Phone Number and Email. Details of the data flow process are provided in Box 1 and Figure 1.

The linkage map for the linked MedicineInsight–NDI dataset included all patients from 239 general practice sites in the MedicineInsight dataset for whom Bloom filters were created. The linkage map contains records for patients who had a linked NDI record (ie, were deceased) and those who had no linked NDI record (ie, assumed to be alive during the entire study period). Patient records that were 'unlinkable', because the minimum number of identifiers required to create the Bloom filters was not present in the patient record, were excluded from the linkage map by CDL.

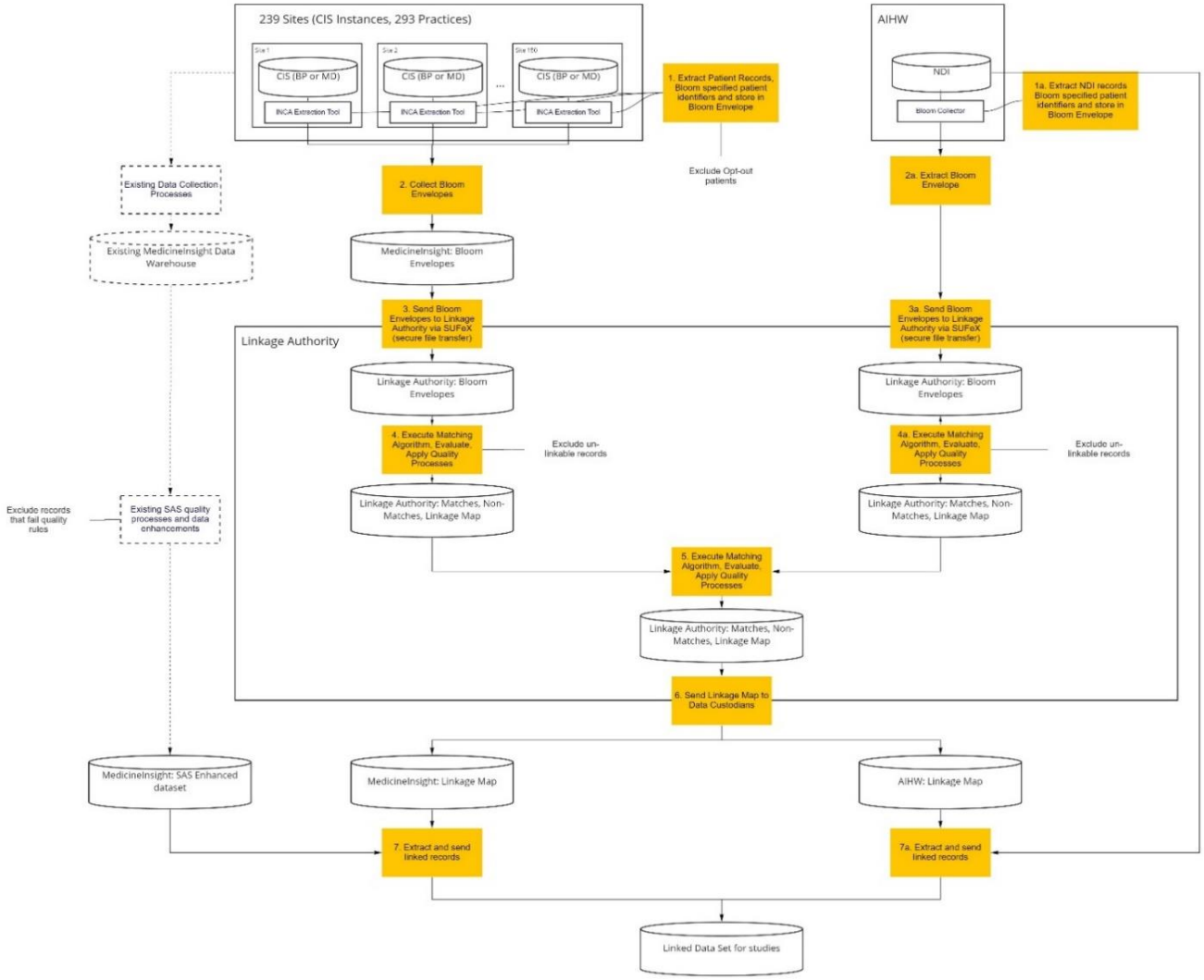
The Bloom filters enabled linkage between MedicineInsight patients and deceased people in the NDI data, as well as linkage between the same patients attending different MedicineInsight practices (duplicate patients between practices), and sometimes the same MedicineInsight practice (duplicate patients within a practice). An exploration of duplicate patients and the impact of their inclusion on the study findings is provided in Section 3.5.

Box 1: Data flow process

The MedicineInsight data (held by NPS MedicineWise) was linked to the NDI data (held by AIHW) by the CDL unit at Curtin University via the following process:

1. CDL ensured the same version of Bloom Processing software was used by NPS MedicineWise and AIHW.
2. CDL provided setup files (ie, project configuration files) to NPS MedicineWise. CDL provided a project-specific encryption key to NPS MedicineWise (Data Warehouse team).
3. NPS MedicineWise set up project configuration files for all MedicineInsight practice sites that use the INCA extraction tool. NPS MedicineWise extracted the relevant patient-level identifiers from the MedicineInsight participating practices using the Bloom Envelope builder and INCA extraction tool (First name, Middle name, Surname, DOB Year, DOB Month, DOB Day, Sex, Address, Suburb, Postcode, Phone Number, and Email). The extracted data (ie, Bloom filters) were encrypted using the project-specific encryption key provided by CDL (ie, hashed using project-specific key). No identifiable information was extracted or held in the MedicineInsight dataset.
4. The Bloom filters produced by the INCA extraction tool and Envelope Builder for each practice were collected and collated by NPS MedicineWise (Data Warehouse team) and provided to CDL with a project-specific MedicineInsight Patient ID, MedicineInsight Site ID (hashed) and the Bloom filter hash. No personally identifiable information or clinical information from MedicineInsight was provided to CDL.
5. CDL set up a project configuration file for the AIHW's NDI dataset. CDL provided a project-specific encryption key to the AIHW.
6. The AIHW executed Bloom Envelope to generate Bloom filters for the NDI dataset. AIHW uses the same Bloom Envelope program and personal identifiers in the NDI data to create Bloom filters, by applying the same Bloom filter key that NPS MedicineWise used to encrypt the MedicineInsight cohort records.
7. AIHW sent the Bloom filter encrypted NDI records (including an AIHW ID) to CDL. No personally identifiable information or content data from NDI was provided to CDL.
8. CDL conducted the linkage using Bloom filter-encrypted cohort records from MedicineInsight and Bloom filter-encrypted personal identifiers from the NDI data ie, privacy preserving linkage methodology linking encrypted cohort records to encrypted NDI files (see Appendix 1).
9. Linkage results: CDL provided a linkage map to the NPS MedicineWise Data Warehouse team. The linkage map contained four variables: Project-specific NPS MedicineWise person ID; MedicineInsight Patient IDs; MedicineInsight Site IDs; and Death flag (0 for patients with no linked NDI record and 1 for patients with a death recorded in NDI). Patients with a death flag of 0 were considered alive during the entire study period as they did not have a death record in the NDI.
10. Linkage results: CDL provided a linkage map to the AIHW team, which contains two variables: Project-specific NPS MedicineWise person ID; and AIHW IDs.
11. Using the linkage map, the NPS MedicineWise Data Warehouse and Health Analytics teams linked the variables of the linkage map with the MedicineInsight patient-level content data and transferred this to NPS MedicineWise researchers: the NPS MedicineWise Health Analytics team securely transferred the MedicineInsight content data for the two categories: a) Death flag 0 and b) "Death flag 1" of the entire cohort to a secure folder separate to the MedicineInsight database, excluding the MedicineInsight ID but including the Project-specific NPS MedicineWise person ID to allow for linkage to the NDI content data.
12. Transfer of content data to NPS MedicineWise researchers by AIHW: the AIHW extracted content data for linked NDI records and securely transferred extracted content data to the named researchers at NPS MedicineWise (to be stored in the secure folder separately to the MedicineInsight database). The content data contained four variables: Project-specific NPS MedicineWise person ID; NDI_year; NDI_date_of_death and NDI_state.
13. Analysis of data was undertaken by NPS MedicineWise's named researchers from the Real World Research team. The Project-specific NPS MedicineWise person ID present in the linkage map was used to merge the MedicineInsight content data with the NDI content data for analyses.

FIGURE 1. DIAGRAMMATIC REPRESENTATION OF THE DATA FLOWS AND PROCESS FOR LINKAGE OF MEDICINEINSIGHT PATIENTS WITH NDI RECORDS BY THE CURTIN DATA LINKAGE UNIT (LINKAGE AUTHORITY).



AIHW = Australian Institute of Health and Welfare; BP = Best Practice; CIS: Clinical Information System; INCA = Precedence Health Care's INCA; MD = Medical Director; NDI = National Death Index; SAS = Statistical Analysis System software

2.4. Study period

The analysis was conducted separately for five patient cohorts in consecutive 2-year time periods: 2011–12; 2013–14; 2015–16; 2017–18; and 2019–20. These study periods were chosen to assess whether the validity of death recording in MedicineInsight has changed over time and with the recency of death, and to understand the accuracy of data from epochs. This information could help inform the selection of cohorts for future studies. The analysis was also conducted for the total study period, 2011 to 2020, by including the results for all patients present in at least one, and up to all, of the five cohorts.

MedicineInsight is a longitudinal open cohort with patients joining and leaving at different time points, with no quality marker of when participants are “lost to follow-up”. It is therefore not appropriate to assess outcomes long after the last encounter for a patient. As such, studies using MedicineInsight data often analyse data for patients who have attended the MedicineInsight practice over a 1- or 2-year period and assess outcomes for these patients during this defined period of attendance at the practice. When validating deaths in MedicineInsight we would only expect deaths to be recorded for patients during the time they are under the care of the MedicineInsight practice. Therefore, for each 2-year cohort of patients attending a MedicineInsight practice, we restricted our analysis to deaths identified as occurring during that 2-year period based on the algorithm for date of death. NDI-recorded deaths occurring between 1 January 2010 (earliest NDI record available) and the beginning of the 2-year study period of interest were also included in our analysis (and quantified separately), thereby capturing the unlikely event that a patient with a clinical encounter recorded in MedicineInsight during the 2-year study period of interest was deceased prior to that study period. Deaths occurring in NDI prior to the study period of interest may also indicate ‘false links’ whereby the PPRL has incorrectly linked two different patients as being the same patient. To account for delays in reporting of deaths to GPs, and the fact that often only year of death is provided in MedicineInsight, for those patients with death recorded in the NDI data during the 2-year period of interest, but not in MedicineInsight, the time period for identifying deaths in the MedicineInsight data was extended to include the 1-year period after the time period of interest. The relevant study periods for this analysis are shown in Table 1.

TABLE 1: THE FIVE CONSECUTIVE STUDY PERIODS AND TIME WINDOWS FOR IDENTIFYING THE LINKED STUDY POPULATION AND DEATHS

Study period	Linked study population (at least 1 clinical encounter identified in MedicineInsight in this period)	Time period for identifying deaths in NDI data (the 'gold standard') †	Time period for identifying deaths in MedicineInsight data	Extended time period for identifying deaths in MedicineInsight data for those patients with death recorded in NDI data*
2011–12	1 Jan 2011 – 31 Dec 2012	1 Jan 2010 (earliest date available) – 31 Dec 2012	1 Jan 2011 – 31 Dec 2012	1 Jan 2011 – 31 Dec 2013
2013–14	1 Jan 2013 – 31 Dec 2014	1 Jan 2010 – 31 Dec 2014	1 Jan 2013 – 31 Dec 2014	1 Jan 2013 – 31 Dec 2015
2015–16	1 Jan 2015 – 31 Dec 2016	1 Jan 2010 – 31 Dec 2016	1 Jan 2015 – 31 Dec 2016	1 Jan 2015 – 31 Dec 2017
2017–18	1 Jan 2017 – 31 Dec 2018	1 Jan 2010 – 31 Dec 2018	1 Jan 2017 – 31 Dec 2018	1 Jan 2017 – 31 Dec 2019
2019–20	1 Jan 2019 – 31 Dec 2020	1 Jan 2010 – 31 Dec 2020	1 Jan 2019 – 31 Dec 2020	1 Jan 2019 – 31 July 2021 (latest date available)

† The time period for identifying deaths in NDI data for each MedicineInsight patient cohort began at the earliest NDI record available (1 January 2010) until the end of the study period of interest, thereby capturing the unlikely event that a patient with a clinical encounter recorded during the 2-year study period of interest was actually deceased prior to that study period.

* To account for delays in reporting of deaths to GPs, and the fact that only year of death is provided in MedicineInsight, the time period for assessing concordance with NDI deaths in the MedicineInsight data is extended to include the 1-year period after the time period of interest.

2.5. Study cohort

The **linked MedicineInsight–NDI study population** included patients who met the following inclusion criteria in the time period of interest (Table 1):

- Included in the linked MedicineInsight–NDI dataset and the August 2021 MedicineInsight data download (238 practice sites in the August download from 239 sites in the linked MedicineInsight–NDI dataset).
- Visited a practice site that had a complete data extract and met specific MedicineInsight data quality requirements (195 quality practice sites from 238 practice sites in the linked MedicineInsight–NDI dataset and the August 2021 download).

- Had valid information for age (0–112 years* in the first year of the 2-year cohort) and sex (male, female, or intersex/indeterminate but not missing).
- Had a CIS status of active, inactive, visitor or deceased. Patients whose CIS status was emergency contact or next of kin were excluded.
- Had at least one clinical encounter during the study period of interest (eg, 2011–12: had at least one clinical encounter between 1 January 2011 to 31 December 2012).
- Not marked as deceased in the MedicineInsight dataset prior to the study period of interest (eg, 2011–12; not marked as deceased prior to 1 January 2011 using the MedicineInsight algorithm for date of death).

The **linked regular attender sub-population** included patients from the linked study population who met the following inclusion criteria in the time period of interest (Table 1):

- Had at least three clinical encounters during the 2-year study period of interest (eg, 2019–20: had at least three clinical encounters between 1 January 2019 to 31 December 2020).

The **linked infrequent attender sub-population** included patients from the linked study population who met the following inclusion criteria in the time period of interest (Table 1):

- Had one to two clinical encounters during the 2-year study period of interest (eg, 2019–20: had one or two clinical encounters between 1 January 2019 to 31 December 2020).

2.6. Definitions

Clinical encounter

A clinical encounter, or any professional exchange between a patient and a healthcare professional (GP or nurse), will be defined as all those encounters at the practice site that are: a) not identified as administrator entries nor encounters that have been transferred/imported from another practice, and b) are not identified by pre-defined 'administration-type' terms found in the 'reason for encounter' field such as "administrative reasons", "forms", and "recall".

Algorithm for fact of death

We examined the accuracy of a novel algorithm for defining fact of death based on information recorded in the relevant fields available to MedicineInsight – patient status, year of death, reason for encounter and diagnosis – as described in Table 2, against the NDI data. The original MedicineInsight "deceased indicator" is a derived variable available to researchers using MedicineInsight data. This indicator flags a patient as deceased if the 'patient status' is recorded as 'deceased' or there is a 'year of death' recorded (even if patient status is not recorded as 'deceased'). However, further investigation of death recording in MedicineInsight found that a significant number of additional deaths (around 5%) are recorded in the reason for encounter field or diagnosis (medical history) field.

* For the 2011–2012 cohort, patients born before 1899 and those born after 2012 were excluded; for the 2013–2014 cohort patients born before 1901 and those born after 2014 were excluded; etc.

The updated NPS MedicineWise algorithm for fact of death (Table 2) was used in combination with the algorithm for date of death (see Table 3) to define fact of death for each time period.

TABLE 2: ALGORITHMS FOR FACT OF DEATH

Algorithm	Definition
Original "Deceased indicator"	Patients were flagged as deceased if: (a) the CIS Patient Status was recorded as 'D' (deceased) or (b) there was a 'Year of Death' recorded.
Updated algorithm for fact of death	Patients were flagged as deceased if: (a) the CIS Patient Status was recorded as 'D' (deceased) or (b) there was a 'Year of Death' recorded or (c) death was recorded in the 'diagnosis' or 'reason for encounter' fields. Patients were flagged as having a recorded death if they had a relevant coded or free text entry in the 'Diagnosis reason' field or the 'Reason for encounter' field. Relevant terms used to identify death included: death, dead, died, deceased, coroner, cremation, fatal, fatality, homicide, killed, life extinct, murder, manslaughter, post-mortem, suicide, 'cause and mortality'. Records identified by a free text string alone were not automatically flagged but individually reviewed to determine whether the text string refers to the event of death in another context (eg, 'partner died', 'suicide attempt'). Sudden infant death syndrome (SIDS) was only included if recorded for patients aged 0–2 years (SIDS occurs in infants less than 1 year of age, however MedicineInsight extracts only year, not month, of birth, so the potential age range was wider). Terms related to fetal death in utero were not included as all such records were for adult patients of childbearing age. The term suicide appears to be used interchangeably with suicide attempt for a small number of patients. Patients could be considered deceased if their records of suicide were followed by records of a coroner report or report to police. Patients with clinical information or diagnoses recorded at encounters after the date of the suicide record were not considered deceased.

Algorithm for date of death

Estimating the date of death in MedicineInsight data is challenging as MedicineInsight only extracts 'year of death' from the CIS and not month or day. For around one-fifth of patients marked as deceased, 'year of death' is missing. Because missing dates in the INCA extraction system are recorded as '1 January 1900', any year-of-death recorded as '1900' is probably invalid. The estimated date of death in MedicineInsight, which was inferred according to the algorithm described in Table 3, was compared with the 'gold standard' date of death from the NDI data. The inferred date of death algorithm prioritised free text entries, including the full date of death recorded in the diagnosis and reason for encounter fields, followed by the date when death was recorded in the diagnosis and reason for encounter fields, over 'year of death' (without day and month) recorded in the Patient table (Table 3). This novel algorithm for date of death was developed by the team of analysts at NPS MedicineWise based on exploration of the data available in the GP CIS, noting there were no internal or external Australian reference sources to guide the development of this algorithm. Post-hoc analysis could be used to refine this algorithm to improve concordance with NDI data.

TABLE 3: ALGORITHM FOR ESTIMATING DATE OF DEATH IN MEDICINEINSIGHT

Method (in this order)	Criteria	Inferred date of death (in priority order)	Quality check
1.	[CIS patient status is recorded as Deceased AND / OR the YOD is present] AND Death is recorded in the 'diagnosis' or 'reason for encounter' fields.	[Date of death recorded as free text in the 'diagnosis' or 'reason for encounter' fields (if date of death is present in both diagnosis and reason for encounter fields choose the earlier of the two)] OR ['diagnosis date' (or record 'created date' if missing 'diagnosis date') where death is recorded in the 'diagnosis' field OR 'visit date' where death is recorded in the 'reason for encounter' field. If both 'diagnosis date' and 'visit date' are available, use the earlier of the two.]	Check concordance of inferred date of death (method 1) with recorded YOD where present. The inferred date of death (method 1) will be used if it is: • concordant with recorded YOD • not concordant with YOD but the date of death was recorded as free text in the 'diagnosis' or 'reason for encounter' fields • before the recorded YOD. The inferred date of death (method 1) will NOT be used if it is: • after the recorded YOD and the date of death was NOT recorded as free text in the 'diagnosis' or 'reason for encounter' fields. In this case methods 3 or 4 will be used to infer date of death.
2.	CIS patient status is NOT recorded as Deceased AND the YOD is missing AND Death is recorded in the 'diagnosis' or 'reason for encounter' fields.	[Date of death recorded as free text in the 'diagnosis' or 'reason for encounter' fields (if date of death is present in both diagnosis and reason for encounter fields choose the earlier of the two)] OR ['diagnosis date' (or record 'created date' if missing 'diagnosis date') where death is recorded in the 'diagnosis' field OR 'visit date' where death is recorded in the 'reason for encounter' field. If both 'diagnosis date' and 'visit date' are available, use the earlier of the two.]	Not applicable – the inferred date of death (method 2) applies
3.	[CIS patient status is recorded as Deceased AND / OR the YOD is present] AND Death is NOT recorded in the 'diagnosis' or 'reason for encounter' fields. AND valid* PATIENT_MODIFIED_DATE is present.	The valid* PATIENT_MODIFIED_DATE unless the YOD (if available) is before the year of the PATIENT_MODIFIED_DATE	Check concordance of inferred date of death (method 3) with recorded YOD where present. The inferred date of death (method 3) will be used if it is: • concordant with recorded YOD • before the recorded YOD and CIS patient status is 'Deceased' The inferred date of death (method 3) will NOT be used if it is: • after the recorded YOD and CIS patient status is 'Deceased' • before or after the recorded YOD and CIS patient status is NOT 'Deceased'. In this case method 4 will be used to infer date of death.

Method (in this order)	Criteria	Inferred date of death (in priority order)	Quality check
4.	[CIS patient status is recorded as Deceased AND / OR the YOD is present] AND Death is NOT recorded in the 'diagnosis' or 'reason for encounter' fields. AND valid* PATIENT_MODIFIED_DATE is missing	Date of the patient's last clinical encounter	Check concordance of inferred date of death (method 4) with recorded YOD where present. The inferred date of death (method 4) will be used if it is: • concordant with recorded YOD The inferred date of death (method 4) will NOT be used if it is: • before or after the recorded YOD In this case method 5 will be used to infer date of death.
5.	Didn't satisfy the criteria or quality checks for Methods 1 to 4	The date of the last diagnosis record or last clinical encounter in the recorded YOD. If both last diagnosis or last encounter are available in the YOD, use the later of the two. OR Approximate date of death as 30/06/YOD	Not applicable – the inferred year of death (method 5) applies.

* '01 JAN 1900' indicates the date is missing (INCA extraction tool only).

CIS = clinical information system; YOD = year of death

2.7. Data analysis and reporting

1. For each of the five time periods of interest (2011–12, 2013–14, 2015–16, 2017–18, 2019–20) the analyst defined the linked study population in MedicineInsight (Table 1) merged with the linked NDI content data. Each of these 2-year cohorts was split into two mutually exclusive sub-cohorts – regular attenders and infrequent attenders. The number of practice sites included in each cohort was recorded. Patients may be included in more than one cohort.

3. For each 2-year patient cohort, the fact of death and date of death were identified:

- using the 'updated deceased indicator' in the MedicineInsight data during that same 2-year period. The method (1–5) for defining date of death was also recorded (Table 3).
- as recorded in the NDI data between 1 January 2010 and the end of the 2-year period.

4. To account for potential delays in reporting of death to the GP, for patients with deaths recorded in NDI data, but not in MedicineInsight, we also searched for deaths in MedicineInsight in the year after the 2-year time period of interest using the updated deceased indicator.

5. Analysis for Aim 1 (examining validity of fact of death)

This analysis was conducted separately for five cohorts in consecutive 2-year time periods – 2011–12, 2013–14, 2015–16, 2017–18, 2019–2020 – for the general study population, regular attenders and infrequent attenders. Patients may be included in more than one of the cohorts. Combined results for 2011 to 2020 were also produced by including all patients present in at least one of the five cohorts, their death records (as defined according to points 3 and 4) and their patient status defined as regular attender if they met that definition in at least one of the 2-year periods, or infrequent attender if not.

The percentage of agreement and measures of accuracy (ie, the sensitivity, specificity, positive predictive value, and negative predictive value) were calculated for fact of death with corresponding 95% confidence intervals (95% CI) (see Box 2, below, for definitions of all measures of agreement used in this study). The 95% confidence intervals were adjusted for clustering by practice sites.

6. Analysis for Aim 2 (examining validity of date of death)

For patients with a record of death in both MedicineInsight and NDI data during the time period of interest, the MedicineInsight inferred date of death was reported in comparison with the NDI date of death as: same date; 1–30 days after; 31–60 days after; > 60 days after; 1–30 days before; 31–60 days before; > 60 days before.

Analysis of the data was conducted using SAS version 9.4 (SAS Institute Inc., Cary, NC, USA). Measures included are descriptive statistics, frequencies and proportions as appropriate. Robust standard errors were used to adjust for clustering by practice site when calculating confidence intervals. Robust standard errors generated using the '*proc surveyfreq'* and its '*clusters*' option were used to adjust for cluster-level variability at practice sites and to calculate 95% confidence intervals. If a particular result was only reported in 1–4 patients or practices, this result has been reported as < 5 in order to preserve the privacy of individuals and practices (with the exception of missing variables).

Box 2: Definitions for measures of agreement and accuracy

Percentage of agreement (PoA) is defined as the number of patients (ie, deceased and non-deceased) in the MedicineInsight data that match those in the NDI data, divided by the total number of patients.

Sensitivity is defined as the proportion of patients with death documented in the NDI that have death recorded in the MedicineInsight data.

Specificity is defined as the proportion of patients without death documented in the NDI that do not have death recorded in the MedicineInsight data.

Positive predictive value (PPV) is defined as the proportion of patients with fact of death recorded in the MedicineInsight data that have fact of death documented in the NDI.

Negative predictive value (NPV) is defined as the proportion of patients without fact of death recorded in the MedicineInsight data that do not have fact of death documented in the NDI.

	National Death Index (gold standard)		
MedicineInsight death algorithm	*Death (+)*	*No death (-)*	*Total*
Death (+)	a	b	a + b
No death (-)	c	d	c + d
Total	a + c	b + d	n

Where:

a = Death was recorded in NDI during the 2-year study period of interest (eg, 2019–20) AND death was recorded in MedicineInsight, either during the 2-year study period of interest or, to account for delays in reporting to the GP, in the following year (ie, 2019–21).

b = Death was recorded in MedicineInsight during the 2-year study period of interest (eg, 2019–20) AND Death was NOT recorded in NDI during the 2-year study period of interest or before (ie, any time pre 2019), to account for delays in reporting to the GP.

c = Death was recorded in NDI during the 2-year study period of interest (eg, 2019–20) AND death was NOT recorded in MedicineInsight, either during the 2-year study period of interest or, to account for delays in reporting to the GP, in the following year (ie, 2019–21).

d = Death was NOT recorded in NDI during the 2-year study period of interest (eg, 2019–20) or before (ie, pre 2019) AND death was NOT recorded in MedicineInsight during the 2-year study period of interest.

Calculation of agreement / accuracy:

PoA = [(a+d)/n] x 100
Sensitivity = [a/(a+c)]
Specificity = [d/(b+d)]
PPV = [a/(a+b)]
NPV = [d/(c+d)]
[multiplied by 100 for percentages]

3. RESULTS

Key findings

Fact of death

- In the whole cohort of 3,067,254 patients, there were 73,527 NDI-recorded deaths during or prior to a time period of interest.
- The PoA between MedicineInsight deaths and those in the NDI was excellent across all years and all patients (regular and infrequent attenders) – all PoA were above 99.0%.
- Accuracy for fact of death was mixed, with excellent specificity, PPV and NPV but poor sensitivity.
- The accuracy of the MedicineInsight deceased algorithm was higher among regular than infrequent attenders and did not vary significantly over time from 2011 to 2020.
- For regular attenders (2011 to 2020):
  - agreement on fact of death was excellent (PoA 99% (95% CI: 99% to 100%) and accuracy was mixed: sensitivity 66% (95% CI: 62% to 70%); specificity 100% (95% CI: 100% to 100%); PPV 96% (95% CI: 96% to 97%); and NPV 99% (95% CI: 99% to 99%).
  - 62,031 regular attender patients had a record of death in the NDI data between 2011 and 2020; 40,930 (66.0%) of these deaths were also recorded in MedicineInsight and 21,101 (34.0%) were not.
  - 42,549 regular attender patients had a death recorded in MedicineInsight between 2011 and 2020; 40,930 (96.2%) of these deaths were also recorded in NDI data and 1619 (3.8%) were not.

Date of death

- 43,747 patients had a record of death in both MedicineInsight and NDI data during a time period of interest
- The MedicineInsight inferred date of death was in agreement, within ±30 days, for 74.4% of 43,747 patients with a record of death in both MedicineInsight and NDI data during the 10-year study period (2011 to 2020).
- The accuracy of the MedicineInsight inferred death date algorithm (±30 days) increased moderately over time from 71.6% for the 2011–12 cohort to 77.1% in the 2019–20 cohort.
- Agreement on death date (±30 days) was slightly higher among regular attenders (75.4%) and lower among infrequent attenders (60.4%) over the 10-year study.

3.1. Study cohorts

The **linked MedicineInsight-–NDI study population** cohorts for each 2-year period are presented in Table 4. The number of patients eligible for the five consecutive 2-year study periods ranged from 821,707 (444,696 regular attenders and 377,011 infrequent attenders) in 2011–12 to approximately 1.36 million (789,629 regular attenders and 568,006 infrequent attenders) in 2019–20. The number of eligible general practices ranged from 156 in 2011–12 to 195 in 2019–20. The total (2011–20) linked population included 3.07 million patients (1.69 million regular attenders and 1.38 million infrequent attenders).

Table 4 also includes the number of deaths identified in the NDI and MedicineInsight datasets for each linked population cohort in the relevant time periods, as defined in Table 1.

TABLE 4: PRACTICE SITES, PATIENT COHORTS AND MEDICINEINSIGHT AND NDI DEATH RECORDS IN EACH 2-YEAR PERIOD (NUMBERS)

Category	2011–12	2013–14	2015–16	2017–18	2019–20	Total (2011–20)*
Number of practice sites	156	175	189	194	195	195
Patients excluded (NDI data issues) - no date of death present in NDI data - > 1 date of death present in NDI data	930 < 5	1038 < 5	1214 < 5	1382 < 5	1487 < 5	2850 6
All patients	821,707	951,131	1,146,817	1,328,497	1,357,635	3,067,254
- NDI deaths in time period of interest	11,872	12,960	15,086	16,363	16,135	73,223
- NDI deaths between 2010 and the beginning of the time period of interest	206	340	626	722	873	304
- MI deaths in time period of interest	7,085	7,938	9,719	10,265	9,667	44,513
- Additional MI deaths identified in 1 year post, for patients with death in NDI in time period of interest but not MI	246	242	247	247	177	1159
Regular attenders (N)	444,696	522,499	633,213	749,112	789,629	1,689,983
- NDI deaths (n)	9,243	10,171	11,759	12,781	12,716	62,031
- MI deaths in time period of interest (n)	6,048	6,854	8,408	8,943	8,494	41,512
- MI deaths in 1 year post for patients with death in NDI	216	209	181	213	141	1,037
Infrequent attenders (N)	377,011	428,632	513,604	579,385	568,006	1,377,271
- NDI deaths (n)	2,835	3,129	3,953	4,304	4,292	11,496
- MI deaths in time period of interest (n)	1,037	1,084	1,311	1,322	1,173	3,001
- MI deaths in 1 year post for patients with death in NDI (n)	30	33	66	34	36	122

*The number of patients and deaths in the 'Total (2011–20)' cohort do not add up to the number of patients and deaths in each of the five consecutive 2-year study periods. The 'Total' number is fewer than all cohorts combined because: (a) a patient may be included in more than one study period but is only counted once in the 'total' cohort; (b) patients may change their status as a regular or infrequent attender across different study periods but for the total cohort a patient was counted once as either regular (if recorded as regular in at least one study period) or infrequent (if recorded as infrequent in all study periods); (c) the same NDI death record may be included in more than one study period but is only counted once in the 'total' cohort; (d) patients with a MedicineInsight death record in the 1 year post study period, may be also recorded as deceased in the following study period but were only counted once in the total column.
MI = MedicineInsight; NDI = National Death Index

3.2. Fact of death

The PoA between the MedicineInsight deceased algorithm and the gold standard NDI data is presented in Figure 2 for the most recent linked regular attender sub-population (2019–20) and in Table 5 for all populations in all time periods. The accuracy of the MedicineInsight deceased algorithm compared to the gold standard NDI data is presented in Figure 3 for the most recent linked regular attender sub-population (2019–20) and in Table 6 for all populations in all time periods.

Overall agreement for fact of death was excellent across all years and all patients (regular and infrequent attenders) with all PoA above 99.0% (Figure 2, Table 5). However, accuracy for fact of death was mixed, with excellent specificity, PPV and NPV but poor sensitivity (Figure 3, Table 6).

Over the 10-year study period (2011 to 2020) among all patients (regular and infrequent attenders):

- 73,527 patients had a record of death in the NDI data (2.40% of 3,067,254); 43,747 (59.5%) of these deaths were also recorded in MedicineInsight (ie, sensitivity 59.5%) and 29,780 (40.5%) were not (Table 6).
- 45,672 patients had a death recorded in MedicineInsight (1.49% of 3,067,254); 43,747 (95.8%) of these deaths were also recorded in NDI data (ie, PPV 95.8%) and 1925 (4.2%) were not (Table 6).

Over the 10-year study period (2011 to 2020) among regular attenders:

- 62,031 patients had a record of death in the NDI data (3.67% of 1,689,983); 40,930 (66.0%) of these deaths were also recorded in MedicineInsight and 21,101 (34.0%) were not (Table 6).
    - Sensitivity 0.66 (95% CI: 0.62 to 0.70)
- 42,549 patients had a death recorded in MedicineInsight (2.52% of 1,689,983); 40,930 (96.2%) of these deaths were also recorded in NDI data (ie, PPV 96.2%) and 1619 (3.8%) were not (Table 6).
    - PPV 0.96 (95% CI: 0.96 to 0.97)

Discussion

In summary, the deaths recorded in MedicineInsight could be validated against NDI deaths, with a relatively small number of ‘false positive’ death records resulting in an excellent PPV (95.8% overall and 96.2% for regular attenders). However, it is clear there is underreporting of deaths in MedicineInsight compared with NDI data, with the MedicineInsight deceased algorithm returning a high number of ‘false negative’ death records resulting in a poor sensitivity (59.5% overall and 66.0% for regular attenders).

Across all time periods, sensitivities and PPVs were higher among regular attenders than infrequent attenders. This finding is expected, as regular attenders are more likely to have complete records and be under the care of that practice than a visitor or temporary patient. However, even among regular attenders, the sensitivity of the MedicineInsight deceased algorithm was poor, with the highest

sensitivity, 70%, achieved by the 2015–16 cohort and the 2017–18 cohort. The accuracy of the MedicineInsight deceased algorithm did not vary significantly over time from 2011 to 2020 (Table 6).

A recently published validation study from the UK,[6] found the sensitivity of the death algorithm in the Clinical Practice Research Datalink (CPRD) primary care dataset was high much higher (98.2%) when validated against Office of National Statistics (ONS) death information in 2013. There are several reasons why roughly a third of deaths among regular attenders were not recorded in MedicineInsight, including: the practice was not notified of the death; the practice was notified of the death but didn't record the death in a field extracted by MedicineInsight; or the patient was no longer attending the practice at the time of death. There are systematic differences in the delivery of primary care between Australia and the UK that could explain the poorer sensitivity of death recording in the Australian setting. For example, in the UK, patients can only register with one general practice in their residential catchment zone at any one time. There are also differences in terms of clinical software and incentives for quality recording against indicators in the UK.

Potential explanations for the small proportion (3.8%) of MedicineInsight deaths that weren't recorded in the NDI include: the death occurred outside of Australia, in which case it would not be recorded in the NDI, the death occurred in Australia but a death certificate was not submitted to, or processed by, the NDI (a rare limitation of the NDI data), the death was entered into the CIS in error by the practice, or the MedicineInsight death algorithm incorrectly identified the death.

Despite the poor sensitivity of the MedicineInsight death algorithm the PoA was high because, between 2011 and 2020 only 2.4% of patients died according to NDI data, meaning the large majority of MedicineInsight patients were concordant for death recording.

Conclusions

For studies where death is an important outcome, MedicineInsight cannot be reliably used without linkage to NDI data. Examples of these types of studies include estimating mortality rates or survival among population groups and assessing the association between exposure to a therapeutic product and death.

Both the high PPV and specificity of deaths recorded in MedicineInsight, when compared with NDI records, indicates MedicineInsight could be used for end-of-life studies, which describe the management of patients in the years prior to their death, provided these patients are representative of all deceased patients. To understand representativeness, further analysis on whether the characteristics of deceased patients who have not been identified in MedicineInsight (false negatives) are not systematically different to those correctly identified as deceased, is recommended.

The PoA was excellent, due to the small proportion of patients who die during the usual reporting period of MedicineInsight studies (often 1 to 2 years). This provides reassurance that for most descriptive epidemiological studies, such as studies on the prevalence and incidence of common chronic conditions, MedicineInsight data can be confidently used without reference to NDI data. However, more caution may be required in studies involving aged populations and high-risk conditions (eg, heart failure, severe chronic kidney disease). An examination of the validity of MedicineInsight

algorithms for death among older patients (eg, 70+ years) would help understand the importance of linkage in these situations.

This study only examined the validity of death identification during periods of attendance at MedicineInsight practices, and results should not be generalised to deaths occurring long after a patients last recorded encounter.

FIGURE 2. AGREEMENT BETWEEN THE MEDICINEINSIGHT DECEASED ALGORITHM AND THE GOLD STANDARD NATIONAL DEATH INDEX (NDI) DATA USING INDIVIDUALLY LINKED DATA FOR THE 2019–2020 LINKED COHORT.

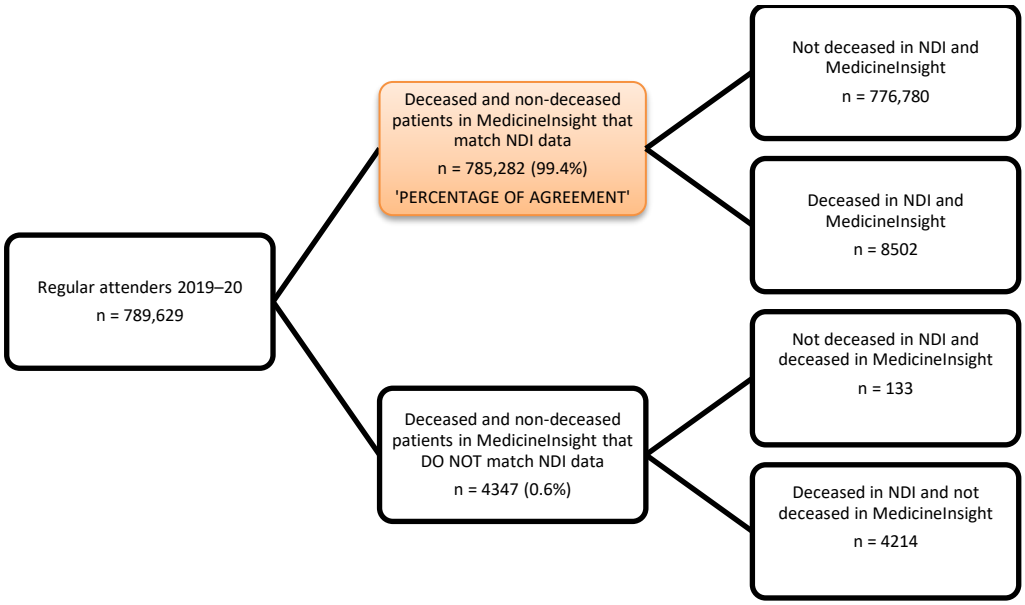
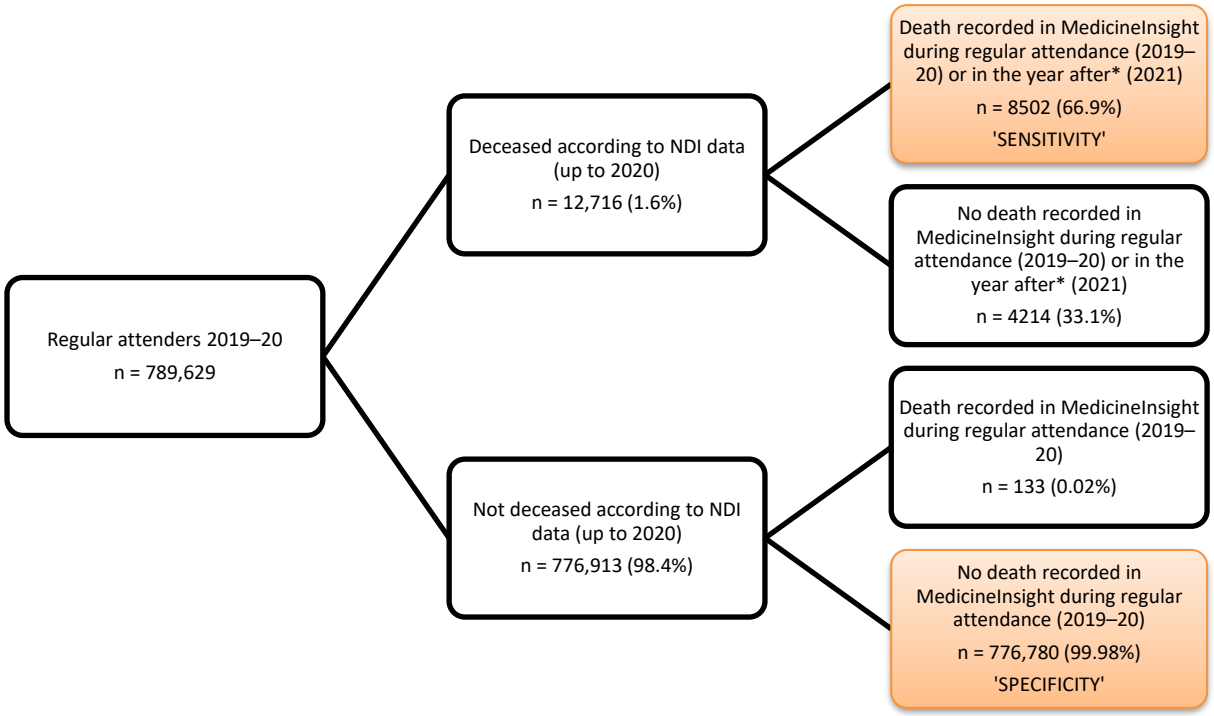
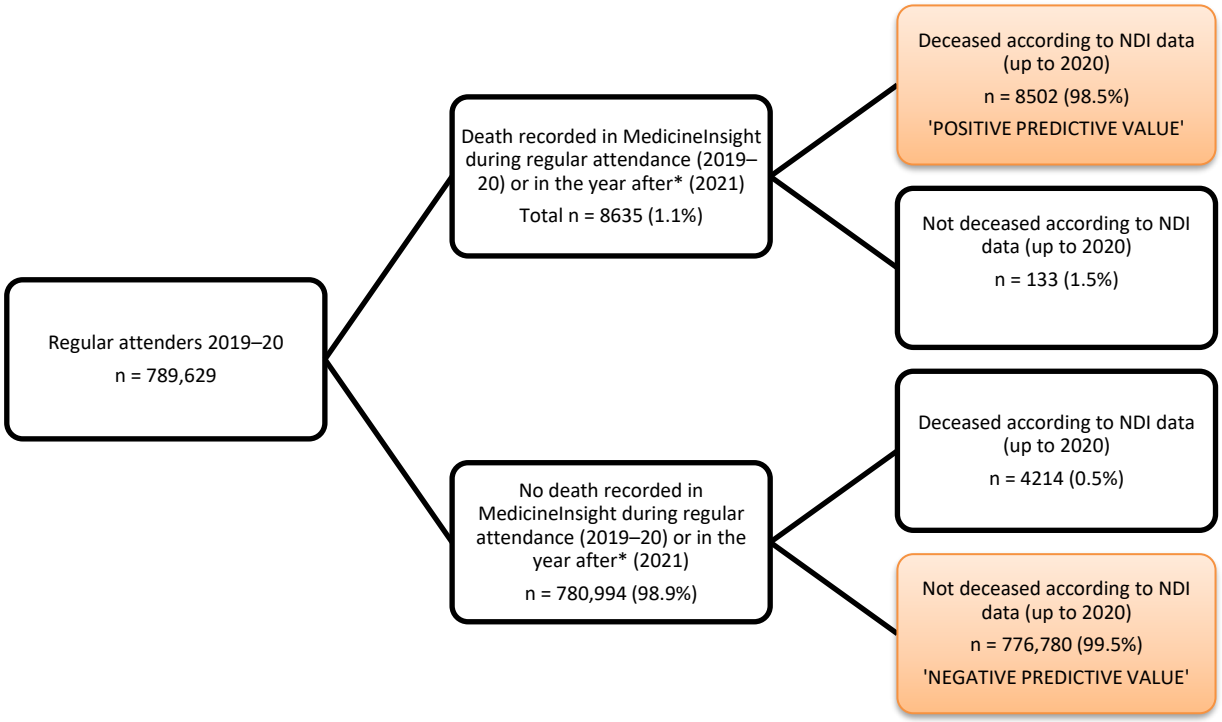


FIGURE 3. ACCURACY OF THE MEDICINEINSIGHT DECEASED ALGORITHM COMPARED TO THE GOLD STANDARD NATIONAL DEATH INDEX (NDI) DATA USING INDIVIDUALLY LINKED DATA FOR THE LINKED REGULAR ATTENDER COHORT DURING 2019–2020.





*Deaths were only identified in the year after the study period for patients who were deceased according to NDI data but not MedicineInsight during the study period

TABLE 5: AGREEMENT BETWEEN MEDICINEINSIGHT AND NATIONAL DEATH INDEX (NDI) DEATH RECORDS (N = 3,067,254)

	All patients		Regular attenders		Infrequent attenders	
Time period	N patients in agreement	Percentage of agreement (95% CI)	N patients in agreement	Percentage of agreement (95% CI)	N patients in agreement	Percentage of agreement (95% CI)
2011–12	816,030	99.3 (99.2, 99.4)	441,019	99.2 (99.0, 99.3)	375,011	99.5 (99.4, 99.5)
2013–14	945,079	99.4 (99.3, 99.5)	518,623	99.3 (99.1, 99.4)	426,456	99.5 (99.4, 99.6)
2015–16	1,140,201	99.4 (99.3, 99.5)	629,399	99.4 (99.3, 99.5)	510,802	99.5 (99.4, 99.5)
2017–18	1,321,220	99.5 (99.4, 99.5)	744,957	99.4 (99.3, 99.6)	576,263	99.5 (99.4, 99.5)
2019–20	1,350,057	99.4 (99.4, 99.5)	785,282	99.4 (99.4, 99.5)	564,775	99.4 (99.4, 99.5)
Total (2011–20)	3,051,734	99.5 (99.4, 99.6)	1,682,117	99.5 (99.5, 99.6)	1,369,617	99.4 (99.4, 99.5)

TABLE 6: ACCURACY OF MEDICINEINSIGHT AND NATIONAL DEATH INDEX FACT OF DEATH 2011–20

	N	NDI+ / MI+ (True+)	NDI- / MI+ (False+)	NDI + / MI- (False–)	NDI- / MI- (True–)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)
	a+b+c+d	a	b	c	d	[a/(a+c)]	[d/(b+d)]	[a/(a+b)]	[d/(c+d)]
2011–12									
All patients	821,707	6,866	465	5,212	809,164	0.57 (0.52–0.62)	1.00 (1.00–1.00)	0.94 (0.92–0.95)	0.99 (0.99–0.99)
Regular patients	444,696	5,915	349	3,328	435,104	0.64 (0.58–0.70)	1.00 (1.00–1.00)	0.94 (0.93–0.96)	0.99 (0.99–0.99)
Infrequent patients	377,011	951	116	1,884	374,060	0.34 (0.29–0.38)	1.00 (1.00–1.00)	0.89 (0.86–0.92)	1.00 (0.99–1.00)
2013–14									
All patients	951,131	7,714	466	5,586	937,365	0.58 (0.53–0.63)	1.00 (1.00–1.00)	0.94 (0.93–0.96)	0.99 (0.99–1.00)
Regular patients	522,499	6,679	384	3,492	511,944	0.66 (0.60–0.71)	1.00 (1.00–1.00)	0.95 (0.93–0.96)	0.99 (0.99–0.99)
Infrequent patients	428,632	1,035	82	2,094	425,421	0.33 (0.29–0.37)	1.00 (1.00–1.00)	0.93 (0.91–0.94)	1.00 (0.99–1.00)
2015–16									
All patients	1,146,817	9,531	435	6,181	1,130,670	0.61 (0.56–0.66)	1.00 (1.00–1.00)	0.96 (0.95–0.97)	0.99 (0.99–1.00)
Regular patients	633,213	8,267	322	3,492	621,132	0.70 (0.65–0.75)	1.00 (1.00–1.00)	0.96 (0.95–0.97)	0.99 (0.99–1.00)
Infrequent patients	513,604	1,264	113	2,689	509,538	0.32 (0.28–0.36)	1.00 (1.00–1.00)	0.92 (0.89–0.95)	0.99 (0.99–1.00)
2017–18									
All patients	1,328,497	10,160	352	6,925	1,311,060	0.59 (0.55–0.64)	1.00 (1.00–1.00)	0.97 (0.96–0.97)	0.99 (0.99–1.00)
Regular patients	749,112	8,891	265	3,890	736,066	0.70 (0.65–0.74)	1.00 (1.00–1.00)	0.97 (0.96–0.98)	0.99 (0.99–1.00)
Infrequent patients	579,385	1269	87	3,035	574,994	0.29 (0.26–0.33)	1.00 (1.00–1.00)	0.94 (0.91–0.96)	0.99 (0.99–1.00)
2019–20									
All patients	1,357,635	9,637	207	7,371	1,340,420	0.57 (0.52–0.61)	1.00 (1.00–1.00)	0.98 (0.97–0.98)	0.99 (0.99–1.00)
Regular patients	789,629	8,502	133	4,214	776,780	0.67 (0.63–0.71)	1.00 (1.00–1.00)	0.98 (0.98–0.99)	0.99 (0.99–1.00)
Infrequent patients	568,006	1,135	74	3,157	563,640	0.26 (0.23–0.30)	1.00 (1.00–1.00)	0.94 (0.91–0.97)	0.99 (0.99–1.00)
Total (2011–20)									
All patients	3,067,254	43,747	1925	29,780	2,991,802	0.60 (0.55–0.64)	1.00 (1.00–1.00)	0.96 (0.95–0.97)	0.99 (0.99–0.99)
Regular patients	1,689,983	40,930	1619	21,101	1,626,333	0.66 (0.62–0.70)	1.00 (1.00–1.00)	0.96 (0.96–0.97)	0.99 (0.99–0.99)
Infrequent patients	1,377,271	2,817	306	8,679	1,365,469	0.25 (0.22–0.27)	1.00 (1.00–1.00)	0.90 (0.88–0.92)	0.99 (0.99–0.99)

MI = MedicineInsight; N = Number; NDI = National Death Index; NPV = Negative Predictive Value; PPV = Positive Predictive Value.

3.3. Date of death

The MedicineInsight algorithm for date of death was reported in comparison with the NDI date of death for 43,747 patients with a record of death in both MedicineInsight and NDI data, during the time period of interest and in total (2011–20) (Table 7). Figure 4 presents the cumulative proportion of patients with death recorded in MedicineInsight and NDI, by difference in days, for the five consecutive 2-year patient cohorts.

Among all patients with death recorded in both MedicineInsight and NDI, the MedicineInsight deaths were in agreement within ±30 days for 74.4% of patients included in the total 10-year study period (2011 to 2020). Agreement on death date over the 10-year study was slightly higher among regular attenders (75.4%) and lower among infrequent attenders (60.4%). The accuracy of the MedicineInsight algorithm for date of death(±30 days) increased moderately over time from 71.6% for the 2011–12 cohort to 77.1% in the 2019–20 cohort.

For just under half of deaths in the total 10-year study period (2011 to 2020) the MedicineInsight inferred death date was 1–30 days after the gold standard NDI date of death, increasing from 38.0% for the 2011–12 cohort to 55.6% for the 2019–20 cohort. Exact (same date) agreement on the date of death was lower, at 17.5% for the total 10-year study period (2011 to 2020), remaining stable over time (Table 7, Figure 4).

Discussion

Despite month and day of death not being extracted by MedicineInsight, in a high proportion of cases the MedicineInsight inferred date of death was on the same date, plus or minus 30 days, as the date of death in NDI (74% for all patients 2011 to 2020, increasing from 71.6% for the 2011–12 cohort to 77.1% in the 2019–20 cohort). However, only 17.5% were on exactly the same day. A recently published validation study on the UK CPRD primary care dataset,[6] in which date of death is extracted or inferred by the patient 'transfer out date', found that 98.8% of CPRD deaths in 2013 were in agreement with ONS mortality data within ±30 days. The exact (same date) agreement on the death date between CPRD and the ONS mortality data was 69.7% across the whole study period (1998 to 2013), increasing from 53.4% in 1998 to 78.0% in 2013.[6]

Conclusions

For studies where timing of death is an important outcome, MedicineInsight cannot be reliably used without linkage to NDI data or equivalent data sources. Examples of these types of studies include estimating mortality rates or survival among population groups and assessing the association between exposure to a therapeutic product and death.

FIGURE 4. DIFFERENCE BETWEEN THE INFERRED DATE OF DEATH IN MEDICINEINSIGHT AND THE NDI DATE OF DEATH AMONG THE FIVE CONSECUTIVE 2-YEAR REGULAR ATTENDER SUB-POPULATIONS (FROM THE 2011–12 COHORT TO THE 2019–20 COHORT) – CUMULATIVE PROPORTION OF PATIENTS WITH A RECORD OF DEATH IN BOTH MEDICINEINSIGHT AND NDI DATA BY DIFFERENCE IN DAYS

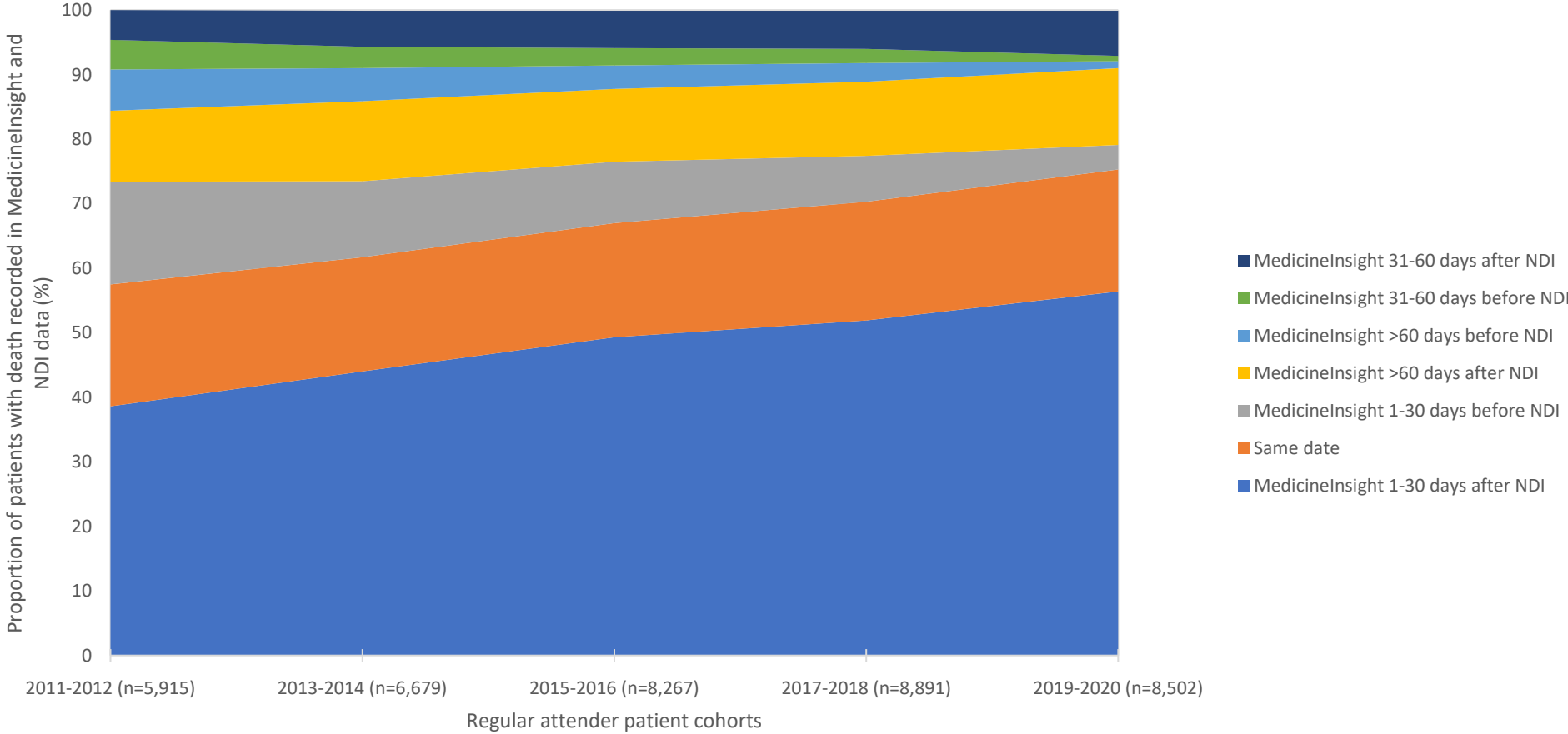


TABLE 7: DIFFERENCE BETWEEN THE INFERRED DATE OF DEATH IN MEDICINEINSIGHT AND THE NDI DATE OF DEATH FOR ALL PATIENTS (REGULAR AND INFREQUENT ATTENDERS) WITH DEATH RECORDED IN BOTH MEDICINEINSIGHT AND NDI FOR THE FIVE CONSECUTIVE 2-YEAR COHORTS (2011 TO 2020)

Difference in date of death	2011–12 n=6866			2013–14 n=7714			2015–16 n=9531			2017–18 n=10,160			2019–20 n=9637			Total (2011–20) n=43,747		
	All	Reg	Infreq	All	Reg	Infreq	All	Reg	Infreq	All	Reg	Infreq	All	Reg	Infreq	All	Reg	Infreq
	n (%)			n (%)			n (%)			n (%)			n (%)			n (%)		
MI > 60 days before NDI	507 (7.4)	375 (6.4)	132 (13.9)	452 (5.9)	342 (5.1)	110 (10.6)	394 (4.1)	299 (3.6)	95 (7.5)	336 (3.3)	258 (2.9)	78 (6.1)	128 (1.3)	98 (1.1)	30 (2.7)	1817 (4.2)	1535 (3.8)	282 (10.0)
MI 31–60 days before NDI	326 (4.7)	272 (4.6)	54 (5.7)	271 (3.5)	222 (3.3)	49 (4.7)	253 (2.7)	225 (2.7)	28 (2.2)	222 (2.2)	195 (2.2)	27 (2.1)	85 (0.9)	70 (0.8)	15 (1.3)	1157 (2.6)	1052 (2.6)	105 (3.7)
MI 1–30 days before NDI	1062 (15.5)	942 (15.9)	120 (12.6)	895 (11.6)	789 (11.8)	106 (10.2)	862 (9.0)	783 (9.5)	79 (6.3)	680 (6.7)	633 (7.1)	47 (3.7)	346 (3.6)	324 (3.8)	22 (1.9)	3845 (8.8)	3602 (8.8)	243 (8.6)
Same date	1240 (18.1)	1116 (18.9)	124 (13.0)	1327 (17.2)	1182 (17.7)	145 (14.0)	1602 (16.8)	1466 (17.7)	136 (10.8)	1755 (17.3)	1636 (18.4)	119 (9.4)	1728 (17.9)	1605 (18.9)	123 (10.8)	7652 (17.5)	7313 (17.9)	339 (12.0)
MI 1–30 days after NDI	2606 (38.0)	2282 (38.6)	324 (34.1)	3261 (42.3)	2942 (44.0)	319 (30.8)	4616 (48.4)	4075 (49.3)	541 (42.8)	5222 (51.4)	4618 (51.9)	604 (47.6)	5354 (55.6)	4799 (56.4)	555 (48.9)	21,037 (48.1)	19,917 (48.7)	1120 (39.8)
MI 31–60 days after NDI	319 (4.6)	275 (4.6)	44 (4.6)	445 (5.8)	374 (5.6)	71 (6.9)	550 (5.8)	481 (5.8)	69 (5.5)	618 (6.1)	528 (5.9)	90 (7.1)	685 (7.1)	593 (7.0)	92 (8.1)	2605 (6.0)	2437 (6.0)	168 (6.0)
MI 61–365 days after NDI	806 (11.7)	653 (11.0)	153 (16.1)	1063 (13.8)	828 (12.4)	235 (22.7)	1254 (13.2)	938 (11.3)	316 (25.0)	1327 (13.1)	1023 (11.5)	304 (24.0)	1311 (13.6)	1013 (11.9)	298 (26.3)	5634 (12.9)	5074 (12.4)	560 (19.9)

All = all patients in the linked study population; Infreq = infrequent attenders; MI = MedicineInsight; NDI = National Death Index; Reg = regular attenders

3.4. Representativeness of the linked MedicineInsight–NDI study population

To assess whether the linked population was representative of the MedicineInsight general study population, the sociodemographic characteristics of the 3.2 million patients included in the linked MedicineInsight–NDI study population were compared to the 3.9 million patients who weren't eligible for inclusion in the linked population (Table 8). The linked MedicineInsight cohort was representative of the general MedicineInsight population in terms of sex, age-group, state/territory, remoteness and socioeconomic status. However, the average age was one year lower among the linked patients (38.8 years) compared to unlinked patients (39.7 years). This finding, while statistically significant ($p = 0.006$), isn't clinically relevant in terms of age impacting, or biasing, the interpretation of findings from the linked study population.

TABLE 8: SOCIODEMOGRAPHIC CHARACTERISTICS OF THE LINKED AND UNLINKED* MEDICINEINSIGHT GENERAL STUDY POPULATIONS (AUGUST 2021 MEDICINEINSIGHT DOWNLOAD)**

	General study population August download Linked plus unlinked patients		General study population August download Linked patients only		General study population August download Unlinked patients only		T test / χ^2 test *** *p-value*
	N	% (95% CI)	N	% (95% CI)	N	% (95% CI)	
Number of practice sites	424	-	195	-	418	-	n/a
Total patients (N)	7,123,136	100.0	3,200,317	44.9 (38.9, 50.9)	3,922,819	55.1 (49.1, 61.1)	n/a
Sex							
Male	3,330,695	46.8 (46.2, 47.3)	1,489,945	46.6 (45.7, 47.4)	1,840,750	46.9 (46.2, 47.7)	0.662
Female	3,791,293	53.2 (52.7, 53.8)	1,709,930	53.4 (52.6, 54.2)	2,081,363	53.1 (52.3, 53.8)	
Indeterminate	1148	0.0 (0.0, 0.0)	442	0.0 (0.0, 0.0)	706	0.0 (0.0, 0.0)	
Age [mean (SD)]	39.3 (22.9)		38.8 (22.9)		39.7 (22.8)		0.006
Age group							
0–9	756,138	10.6 (10.1, 11.1)	362,471	11.3 (10.6, 12.1)	393,667	10.0 (9.4, 10.7)	0.118
10–19	782,716	11.0 (10.6, 11.4)	356,195	11.1 (10.6, 11.6)	426,521	10.9 (10.3, 11.4)	
20–29	1,048,154	14.7 (14.0, 15.5)	464,089	14.5 (14.0, 15.0)	584,065	14.9 (13.6, 16.2)	
30–39	1,297,365	18.2 (17.5, 18.9)	582,272	18.2 (17.3, 19.1)	715,093	18.2 (17.2, 19.3)	
40–49	974,690	13.7 (13.4, 14.0)	441,218	13.8 (13.4, 14.2)	533,472	13.6 (13.2, 14.0)	
50–59	784,216	11.0 (10.8, 11.3)	349,980	10.9 (10.6, 11.2)	434,236	11.1 (10.7, 11.4)	
60–69	645,787	9.1 (8.8, 9.4)	283,960	8.9 (8.5, 9.2)	361,827	9.2 (8.8, 9.7)	
70–79	457,949	6.4 (6.1, 6.8)	196,409	6.1 (5.7, 6.6)	261,540	6.7 (6.2, 7.1)	
80–89	243,032	3.4 (3.2, 3.6)	104,507	3.3 (2.9, 3.6)	138,525	3.5 (3.2, 3.8)	
90+	133,089	1.9 (1.7, 2.0)	59,216	1.9 (1.6, 2.1)	73,873	1.9 (1.7, 2.1)	
State/territory							
ACT	174,279	2.4 (0.9, 4.0)	41,079	1.3 (0.0, 2.6)	133,200	3.4 (0.8, 6.0)	0.819
NSW	2,560,114	35.9 (30.3, 41.6)	1,086,407	33.9 (25.3, 42.6)	1,473,707	37.6 (30.1, 45.0)	

	General study population August download Linked plus unlinked patients		General study population August download Linked patients only		General study population August download Unlinked patients only		T test / χ^2 test *** *p-value*
NT	145,707	2.0 (0.3, 3.8)	84,484	2.6 (0.0, 5.5)	61,223	1.6 (0.0, 3.7)	
QLD	1,485,347	20.9 (16.5, 25.2)	718,028	22.4 (15.7, 29.2)	767,319	19.6 (13.8, 25.3)	
SA	110,779	1.6 (0.6, 2.5)	30,770	1.0 (0.1, 1.8)	80,009	2.0 (0.4, 3.7)	
TAS	384,705	5.4 (2.9, 7.9)	171,171	5.3 (1.0, 9.7)	213,534	5.4 (2.5, 8.4)	
VIC	1,353,022	19.0 (14.5, 23.5)	612,410	19.1 (12.3, 26.0)	740,612	18.9 (13.0, 24.8)	
WA	909,183	12.8 (8.6, 16.9)	455,968	14.2 (8.2, 20.3)	453,215	11.6 (5.9, 17.3)	
Remoteness							
Major city	1,506,310	21.1 (17.2, 25.1)	615,518	19.2 (13.6, 24.9)	890,792	22.7 (17.3, 28.1)	
Inner regional	4,714,094	66.2 (61.3, 71.0)	2,160,864	67.5 (60.2, 74.9)	2,553,230	65.1 (58.6, 71.5)	0.793
Outer regional	778,386	10.9 (8.2, 13.7)	361,465	11.3 (7.1, 15.5)	416,921	10.6 (6.9, 14.3)	
Remote/very remote	124,346	1.7 (0.9, 2.6)	62,470	2.0 (0.7, 3.2)	61,876	1.6 (0.5, 2.7)	
Socioeconomic status							
1 (most disadvantaged)	1,007,818	14.1 (11.8, 16.5)	427,232	13.3 (9.8, 16.9)	580,586	14.8 (11.7, 17.9)	
2	1,224,420	17.2 (14.5, 19.8)	481,468	15.0 (11.8, 18.3)	742,952	18.9 (15.0, 22.9)	
3	1,538,648	21.6 (18.8, 24.4)	811,959	25.4 (20.5, 30.3)	726,689	18.5 (15.5, 21.5)	0.145
4	1,498,201	21.0 (18.3, 23.8)	688,512	21.5 (17.5, 25.6)	809,689	20.6 (17.0, 24.3)	
5 (most advantaged)	1,854,049	26.0 (22.3, 29.7)	791,146	24.7 (19.8, 29.6)	1,062,903	27.1 (21.8, 32.4)	

*Unlinked population includes patients from the GRHANITE sites (81% of all unlinked population) and patients from INCA sites (19% of all unlinked population) that were not linked to NDI death data; **Patients with at least one clinical encounter between 2010–20, valid age in 2010 (not older than 112), no death recorded prior to 2010; ***Statistical tests of socio-demographic characteristics for linked and unlinked populations, ie, two sample tests.

3.5. Exploratory analysis to identify duplicate patients

The PPRL using “Bloom filters” enabled linkage between MedicineInsight patients and deceased people in the NDI data, as well as linkage between the same patients attending different MedicineInsight practices (duplicate patients between practices) and the same MedicineInsight practice (duplicate patients within a practice).

Among the 789,629 regular attender patients in the most recent 2-year linked MedicineInsight–NDI study cohort (2019–20), who had at least three clinical encounters from 1 January 2019 to 31 December 2020, 769,156 (97.4%) were considered unique patients and 20,473 (2.6%) were identified as duplicate patients, matched to either one other patient (2.5%) or more than one other patient (0.1%) (Figure 5). Most duplicate patients (n = 19,814) were only matched to one other patient, and of these, 5% were identified as the same patient within practice sites and 95% were identified as the same patient between practice sites (Figure 5). Among all 1.4 million patients in the 2019–20 linked study cohort including both infrequent and regular attenders, the proportion of duplicate patients was higher (6.0%) (Figure 6) then for regular attenders only (Figure 5).

FIGURE 5: FLOWCHART FOR THE 2019–20 LINKED MEDICINEINSIGHT–NDI STUDY POPULATION – REGULAR PATIENTS DUPLICATE ANALYSIS

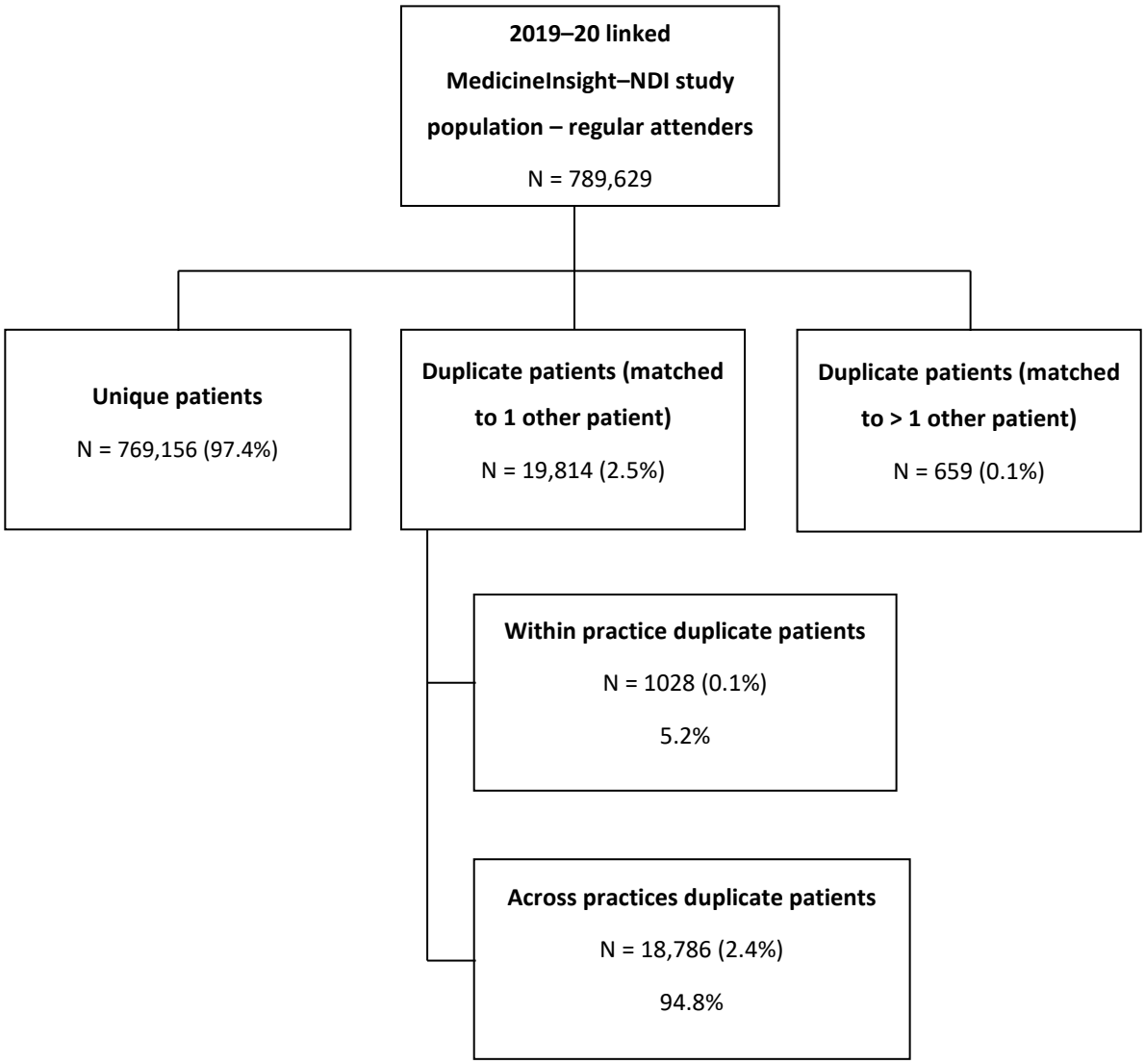
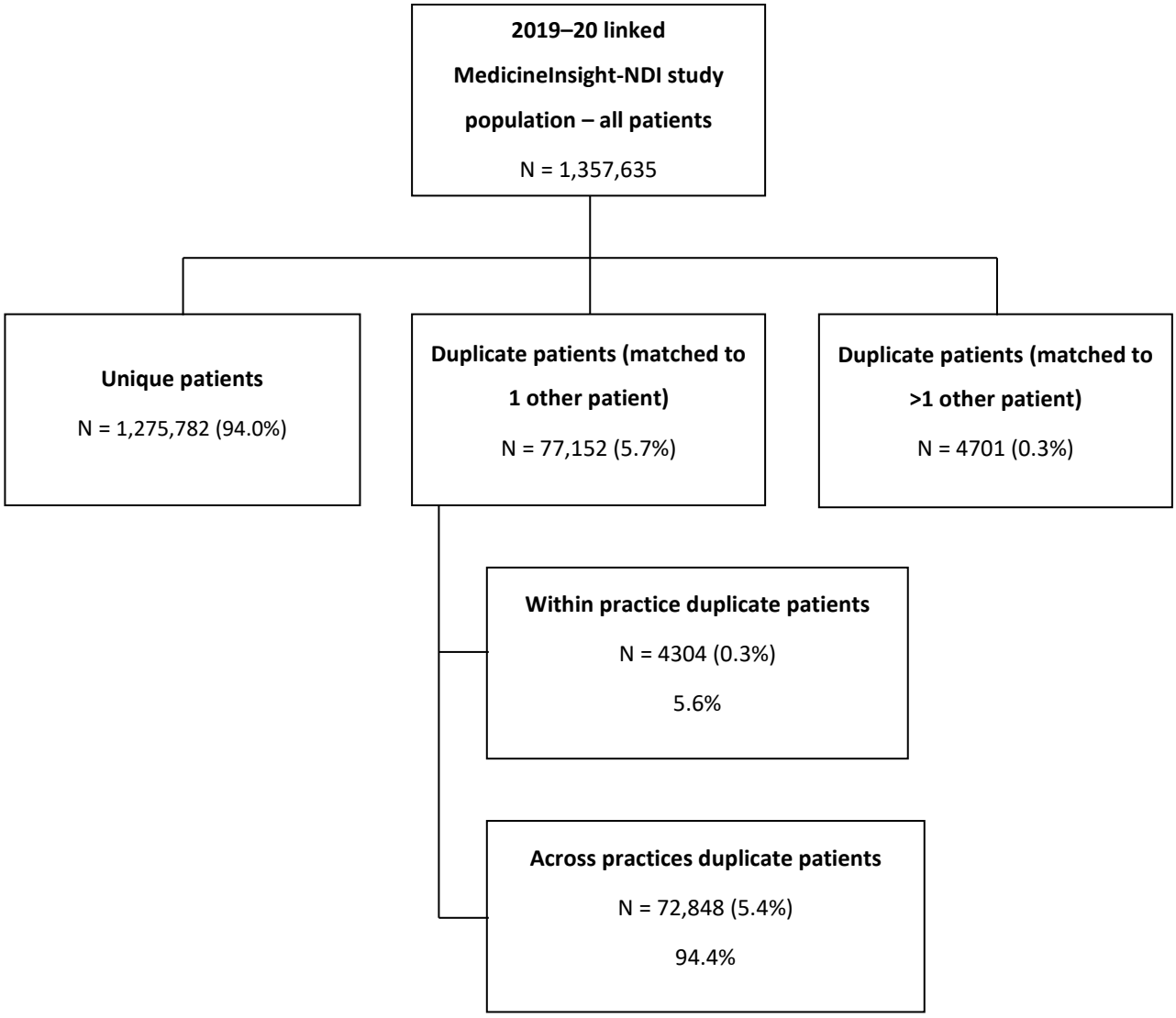


FIGURE 6: FLOWCHART FOR THE 2019–20 LINKED MEDICINEINSIGHT–NDI STUDY POPULATION – ALL PATIENTS DUPLICATE ANALYSIS



Sensitivity analysis

A post-hoc analysis excluding duplicate patients was conducted for the 2019–20 all patients and regular attender populations, however the results for agreement and accuracy for fact of death did not change substantially (data not shown).

Preliminary quality assessment of the linkage

As a preliminary assessment of the quality of the privacy preserving linkage, we measured the concordance (or agreement) of a selection of patient characteristics between patients who were matched to at least one other patient (Table 9). There were 9907 pairs of patients who were matched to one other patient and, of these, 98.3% were concordant for year of birth, 99.4% concordant for gender and 89.9% concordant for the presence or absence of a diagnosis of hypertension. Lower concordance was observed for postcode at 71.4%. This is not surprising, as it is likely that patients who visit more than one practice may have changed addresses throughout the year.

TABLE 9: PRELIMINARY QUALITY ASSESSMENT OF THE BLOOM FILTER LINKAGE MATCHES; CONCORDANCE OF KEY PATIENT CHARACTERISTICS AMONG 9907 DUPLICATE PATIENT PAIRS IN REGULAR STUDY COHORT 2019–20

Patient characteristic	Duplicate patient pairs with matching records (n = 9.907)
	Concordance, %
Year of birth	98.3
Gender	99.4
Postcode*	71.4
Hypertension	89.9

*Excluding 28 pairs where at least one patient had a missing postcode.

Discussion

The proportion of regular attender patients identified as duplicate patients in 2019–20 (2.6%) was lower than estimates reported in previous MedicineInsight studies[3,9] (3.0% to 3.8%), however, the proportion of all patients (regular and infrequent attenders) identified as duplicates in 2019–20 (5.9%) was higher than previous estimates. Regardless of the population assessed, the inclusion of duplicate patients in these analyses did not impact the study findings.

The excellent agreement between patient characteristics of duplicate patient matches provides good reassurance of the quality of the linkage matches. However, to comprehensively assess the quality of the privacy preserving linkage, further validation studies are recommended. For example, duplicate patients identified through linkage could be validated as true duplicates ‘at source’ by reidentifying patients back at the practice and checking the EHR.

GUIDE TO INTERPRETING THE DATA

When interpreting the information presented in this report, readers should note the following caveats and/or assumptions related to the study methods:

- MedicineInsight data are dependent on the accuracy and completeness of data recorded in, and available for extraction from, the general practice CIS.
- Identification of deaths is dependent on GPs recording these items in their CIS. Deaths may be under-reported in MedicineInsight data depending on GPs' recording practices.
- The information in this report represents completeness of data recorded in fields accessible to MedicineInsight and may not indicate non-recording of data. It is possible that some GPs may record information about deaths in different places within the CIS, for example in the progress notes (which are not available to MedicineInsight), and this can affect validity estimates in MedicineInsight data.
- Deaths that occur outside of Australia will not be recorded in the gold standard NDI and occasionally death certificates are not submitted to, or processed by, the NDI.
- The PPRL may have produced a small number of 'false links', although results from our preliminary quality assessment indicate good concordance among most linked patients within MedicineInsight.
- The algorithm for identifying deaths in MedicineInsight cannot be replicated in the unlinked MedicineInsight data because it was not independent of the NDI data. Specifically, for those patients with death recorded in the NDI data during the 2-year period of interest, but not in MedicineInsight, the time period for identifying deaths in the MedicineInsight data was extended to include the 1-year period after the time period of interest. This method was applied to best reflect GP workflows, accounting for delays in reporting of deaths to GPs, and the fact that often only year of death is provided in MedicineInsight. In the total cohort (2011–2020) 1159 extra deaths were identified in MedicineInsight in the extended 1-year period, which equates to 2.5% of 45,672 total deaths in MedicineInsight and 1.5% of 73,527 total deaths in NDI (Table 4). Based on these figures, the accuracy of the MedicineInsight algorithm, without modification based on the NDI data, would be slightly worse – there would be a moderate increase in the number of false negatives in MedicineInsight and the sensitivity of the fact of death algorithm would be slightly reduced.
- Deaths occurring in the NDI prior to the study period of interest may indicate 'false links', whereby the PPRL has incorrectly linked two different patients as being the same patient. Alternatively, these links may be correct and indicate that administrative, rather than clinical, encounters were recorded in MedicineInsight for patients who were deceased. Across all study cohorts, NDI deaths prior to the study period of interest occurred in roughly 0.2% to 0.6% of all patients and in roughly 2–3% of all deaths (according to NDI). The inclusion of these patients, regardless of this potential error, would not have significantly impacted study findings. However, as the quality of these linked records is questionable, future studies should consider excluding them.
- This study only examined the validity of death identification during periods of attendance at MedicineInsight practices, and results should not be generalised to deaths occurring long after a patient's last recorded encounter.

REFERENCES

1. Precedence Health Care. INCA. 2020; https://precedencehealthcare.com/inca/. Accessed 15 June 2020.
2. Havard A, Manski-Nankervis JA, Thistlethwaite J, et al. Validity of algorithms for identifying five chronic conditions in MedicineInsight, an Australian national general practice database. *BMC Health Serv Res.* 2021;21(1):551.
3. Busingye D, Myton R, Mina R, Thistlethwaite J, Belcher J, Chidwick K. *MedicineInsight report: Validation of the MedicineInsight database: completeness, generalisability and plausibility.* Sydney: NPS MedicineWise;2020.
4. Ralph SJ, Espinet AJ. Increased All-Cause Mortality by Antipsychotic Drugs: Updated Review and Meta-Analysis in Dementia and General Mental Health Care. *Journal of Alzheimer's disease reports.* 2018;2(1):1-26.
5. Roxburgh A, Hall WD, Dobbins T, et al. Trends in heroin and pharmaceutical opioid overdose deaths in Australia. *Drug and alcohol dependence.* 2017;179:291-298.
6. Gallagher AM, Dedman D, Padmanabhan S, Leufkens HGM, de Vries F. The accuracy of date of death recording in the Clinical Practice Research Datalink GOLD database in England compared with the Office for National Statistics death registrations. *Pharmacoepidemiol Drug Saf.* 2019;28(5):563-569.
7. Bird S. How to complete a death certificate - a guide for GPs. *Aust Fam Physician.* 2011;40(6):446-449.
8. Australian Bureau of Statistics. Cause of death certification. 2008; https://www.ausstats.abs.gov.au/Ausstats/subscriber.nsf/0/475BC02643DB45EDCA25750B000E38A4/$File/1205055001_2008.pdf. Accessed 30 June 2020.
9. NPS MedicineWise. General Practice Insights Report July 2017–June 2018. 2019; https://www.nps.org.au/assets/NPS/pdf/General-Practice-Insights-Report_2017-18.pdf. Accessed 24 June 2020.

APPENDIX 1

What are Bloom filters?

Bloom filters enable privacy preserving linkage by encoding patient identifiers 'at source'. These can be extracted and linked probabilistically to identifiers from other datasets that have been encoded using exactly the same process.

To enable the bloom filter linkage, the 'data extraction tool' (INCA in this study) performs the following steps **within** the general practice:

a) The data is extracted from a compatible 'clinical management system' (eg, Best Practice, Medical Director).

b) The patient-identifying particulars then undergo a complex encoding process, in which a one-way encoding function obscures identifiers while allowing for probabilistic linkage. This process of cryptographically 'encoding at source' is fundamental to 'privacy preserving record linkage'.

The encoding algorithm used for this purpose has been developed by Curtin University and is described with selected excerpts below from JH Boyd, *Record Linkage Techniques: Exploring and developing data matching methods to create national record linkage infrastructure to support population level research*:*

i. Privacy preserving record linkage using Bloom filters works by encoding personally identifying information into a set of 'binary vectors' which is a sequence of 1s, and 0s.
ii. A Bloom filter begins as an array or series of memory locations or 'boxes' each of which holds a single item of data, of a set length, with all elements set to zero.
iii. Each field (eg, first name) is broken down into overlapping sets of letters (referred to as bigrams or n grams) and these are passed through a series of cryptographic 'hash functions'. A hash function is an algorithm which produces a fixed-length output with several important properties. Firstly, given the same input, it will always produce the same output. For example, the same overlapping letters always produce the same hash value. Secondly, the hash function is one way, meaning it is not possible to determine the encoded letters from any given hash value (ie, it is irreversible and therefore privacy preserving).
iv. Different hash passwords can be used to produce different output. These hashes are then computed with respect to the length of the Bloom filter.
v. This process allows us to map the encoded personal information to a position in the Bloom filter. These positions are then set to 1.
vi. Two Bloom filters can be compared to each other using a match score. The dice coefficient results in a score between 0 and 1, where a higher score reflects greater similarity once linked.
vii. The encryption techniques used in privacy preserving linkage with Bloom filters means that probabilistic-type techniques can be used during the matching process. These techniques allow for small errors such as spelling mistakes which greatly improve linkage quality.

Evaluations using real data found linkage quality using Bloom filters to be equivalent to those achieved using unencrypted personal identifiers, and greater than that of other implemented privacy preserving methods.

* Boyd, JH. Record Linkage Techniques: Exploring and developing data matching methods to create national record linkage infrastructure to support population level research. Thesis. Curtin University December 2016. Accessed from: https://espace.curtin.edu.au/bitstream/handle/20.500.11937/54163/Boyd%20James%202017.pdf?isAllowed=y&sequence=1