International Journal of Population Data Science

Journal Website: www.ijpds.org

Using data linkage to monitor COVID-19 vaccination: development of a vaccination linked data repository

Tom Eitelhuber[1,*], Sera Ngeh[1], Lauren Bloomfield[1,2], Bhaval Chandaria[3], and Paul Effler[1]

Submission History

Submitted:	**18/01/2022**
Accepted:	**16/11/2022**
Published:	**15/12/2022**

[1]Western Australian Department of Health, Communicable Disease Control Directorate, Perth, WA
[2]The University of Notre Dame Australia, School of Medicine, Fremantle, WA
[3]Western Australian Department of Health, Information and System Performance Directorate, Perth, WA

Abstract

The COVID-19 Vaccination Linked Data Repository (CVLDR) was established in 2021 to assist with the implementation and management of the COVID-19 vaccination program in the State of Western Australia (WA). The CVLDR contains a number of datasets including the Australian Immunisation Register, hospital admissions, emergency department attendances, notifiable infectious disease, and laboratory data. Datasets in the CVLDR are linked using a probabilistic method at the WA Department of Health. Quality assurance mechanisms have been established to identify and mitigate potential errors in the linkage. Each of the datasets has varying degrees of data quality and completeness, however most are of high standard, underpinned by legislation. The linking of the datasets within the CVLDR has allowed for increased public health utility in the immunisation program including the areas of vaccine safety, effectiveness, and coverage.

Keywords
immunisation; data; linkage; data linkage; vaccination; covid-19; vaccines; vaccine safety; vaccine effectiveness

*Corresponding Author:
Email Address: tom.eitelhuber@health.wa.gov.au (Tom Eitelhuber)

https://doi.org/10.23889/ijpds.v5i4.1730

Introduction

In recent years, the data delivery paradigm of ‘linked data repositories’ has become more common. While traditional linked data extracts were typically tailored for a specific piece of analysis, linked data repositories may host holistic datasets applicable, and able to be utilised by a range of users. Currently in Australia, a number of States and Territories are using data linkage for monitoring of immunisation activities in the COVID-19 response. Globally, the United States Centres for Disease Control have used data linkage for vaccine safety and effectiveness monitoring since 1990 [1] and several other countries are using population-level data linkage for monitoring COVID-19 immunisation program outcomes [2–5]. Such linkage projects allow rapid matching of data such as demographics, vaccination status, pathology results and hospitalisations, which help to inform the pandemic response.

In Western Australia (WA), immunisation programs are managed by the WA Department of Health (WADOH), and are responsible for the planning, management, education, and monitoring of all National Immunisation Program vaccines, as well as seasonal vaccinations. In preparation for the start of the COVID-19 vaccination program, there was a need to include immunisation data to existing linked datasets to improve the program’s capability to monitor and analyse the safety and effectiveness of COVID-19 vaccines. Consequently, the COVID-19 Vaccination Linked Data Repository (CVLDR) was created to assist with the public health management of the COVID-19 immunisation program. This article describes the development of the CVLDR, with a focus on the Australian Immunisation Register (AIR), including its governance, technical design, early usage, and key benefits in WA.

Methods

Overview and purpose of dataset

The CVLDR consists of a number of State-Wide datasets, including data from public and private hospitals and pathology laboratories (Figure 1). The ‘backbone’ of the CVLDR is the AIR, which is the whole of-life immunisation register of all vaccinations received in Australia and overseas. In April 2021, the Commonwealth Department of Health approved the AIR to be linked in all State and Territory Health Departments around Australia, allowing greater public health capabilities to monitor immunisation program activities including;

- monitoring coverage, safety, and effectiveness of vaccines across Australia
- identifying under-vaccination rates in specific areas of Australia that may be at greater risk of disease outbreaks
- immunisation policy and research
- eligibility for certain family assistance payments
- entry to child-care and school
- proof of vaccination for employment or travel purposes

Another dataset that was linked in 2021 was the WA Vaccine Safety Surveillance (WAVSS) system. This is the reference database for analysing and reporting serious adverse events following immunisation (AEFI). The WAVSS system accepts reports from both clinicians and the public, without the prior need to attribute causation [6]. The system is primarily designed to capture and investigate serious and rare adverse events, with referral services to immunology clinics [7]. Post-licensure surveillance of AEFIs is an important component of any immunisation program and is essential for maintaining public and provider confidence.

CVLDR data governance

The access, use, and disclosure of AIR data is governed by the *Australian Immunisation Register Act 2015* (‘the AIR Act’). The AIR Act describes a range of approved functions for which AIR data can be used. Additionally, the AIR Act permits certain use and disclosure of identifiable AIR data by ‘prescribed bodies,’ which include all State and Territory Health Departments. The linkage of the WAVSS dataset was approved on the basis that it is consistent with the collection’s purpose of vaccine safety surveillance.

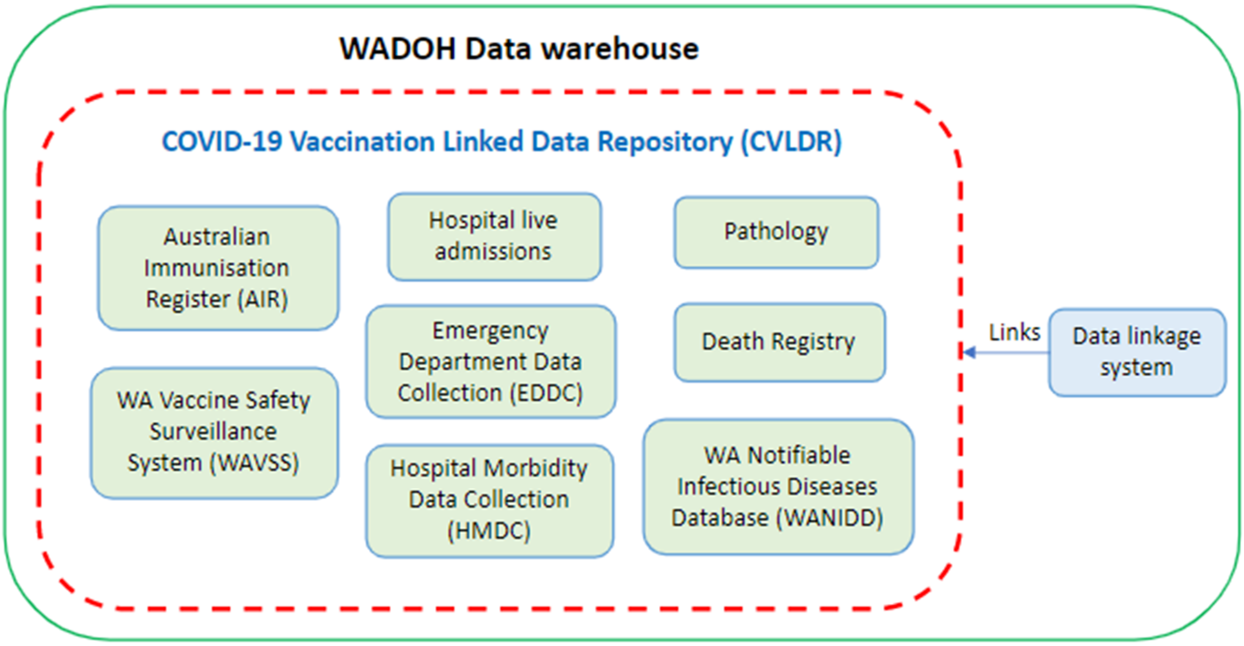
The WADOH is the custodian of the CVLDR, which is stored on a secure, password protected server at the department. The CVLDR can only be accessed through the department’s network by immunisation staff that use the database for their public health work activities. WADOH’s delegated custodianship enables the sharing of AIR data with third parties, provided that the necessary application and approval protocols are met. Similarly, WADOH has the ability to share data from the remaining CVLDR datasets, provided that the purpose for sharing is consistent with relevant legislation and policies. There is significant potential in making these linked datasets available to third parties for research, however the CVLDR itself was approved for WADOH operational use and not intended for direct access outside of this context. Therefore, any requests by third parties would be carried out through pre-existing mechanisms for the release of linked data.

Probabilistic linkage of immunisation data

Customised field matching algorithms are employed to compare the AIR and WAVSS records to a wide range of datasets within the linkage system (not just those of interest for immunisation analysis), to ensure the largest possible yield of correct links [8]. This includes, but is not limited to, birth and death registrations, hospital admissions, and electoral roll data. Different datasets may contain different fields and be subject to different formatting, codification and standards of quality and completeness. These variations inform how the data is strategically processed, cleansed, and linked. In the case of the AIR and WAVSS linkages, data is cleaned and standardised to align with other datasets within the linkage system.

A range of quality assurance mechanisms are also utilised by the Linkage Team to identify and mitigate potential linkage errors, such as chain sampling or duplicate sampling. When a new AIR record tries to link to a chain with an existing AIR record, a dynamic ‘link flagging’ system identifies unlikely or impossible events such as a vaccination occurring after death [9].

Figure 1: Visual of the COVID-19 Vaccination Linked Data Repository (CVLDR)


WADOH Data warehouse
COVID-19 Vaccination Linked Data Repository (CVLDR)
Australian Immunisation Register (AIR)
Hospital live admissions
Pathology
Emergency Department Data Collection (EDDC)
Death Registry
WA Vaccine Safety Surveillance System (WAVSS)
Hospital Morbidity Data Collection (HMDC)
WA Notifiable Infectious Diseases Database (WANIDD)
Links
Data linkage system

Following linkage, the AIR and WAVSS data and associated linkage keys are loaded into the WADOH Data Warehouse, which is a SQL Server database maintained, accessed, and used by WADOH staff. Linkage keys are shared between the datasets, enabling them to be integrated as the CVLDR. At time of writing, the CVLDR contains linked records for over 2.9 million people.

Results

Data quality and completeness of data

The quality of any administrative data impacts the quality and timeliness of its linkage. Incomplete and inconsistent demographic data require additional manual review of potential matches, sometimes resulting in 'missed' links. Work on this project has highlighted the importance of robust data sources. The AIR data is received via the Commonwealth's portal, which required a time-consuming, manual process of downloading and transferring files. Due to the lack of stable and unique record identifiers in the initial available AIR data, this has impacted the completeness of the linkage, however the move to the automated feed is expected to mitigate this issue.

All datasets used within the CVLDR are secondary datasets which have been collected for administrative purposes. As such, the potential for measurement bias exists if variables are recorded incorrectly. This is of particular concern if any variables are not missing at random i.e. there is a higher propensity for missing variables in certain subpopulations. In addition, any analyses conducted using these data collections are only able to control for potential confounders already contained within the CVLDR, and as such there may be some uncontrolled residual confounding. To the best of our knowledge, there are no systematic data recording errors that affect the use of these data collections to make inferences about this population. Specific source of potential bias related to each data collection are detailed below. Table 1 provides a summary of the datasets within the CVLDR, including a brief description and how often the dataset is updated and linked.

Australian Immunisation Register (AIR)

Everyone in Australia who holds a Medicare card or receives a vaccine, should have a record within the AIR. The accuracy and completeness of data reported in AIR have been previously assessed and found to be high [10], and likely improved since a law requiring providers to upload vaccination records to the AIR within 10 days of vaccine administration came into effect on 1 July 2021 [11–13]. At the time of writing, access to the AIR data is being transitioned from downloaded files to an automated feed. This will improve the quality of the AIR, as people without a Medicare number will be able to be included in the data. A small subsection of the population; those with neither a Medicare number or any vaccination records will not be recorded in the AIR.

Emergency Department Data Collection (EDDC)

Only one primary ICD-10 diagnosis can be selected for an attendance at the emergency department. The diagnosis code is often selected before all examinations and testing have been conducted, therefore the diagnosis code does not always correspond to the main reason for attending the hospital. The EDDC dataset however could be used for rapid onset events that have an obvious clinical presentation compared to more complex conditions, or those that present with generalised symptoms. The EDDC data dictionary can be found through this link.[1]

[1] Data Dictionary - EDDC (health.wa.gov.au)

Table 1: Datasets held in the COVID-19 Vaccination Linked Data Repository (CVLDR)

Dataset	Information used	Description	Update frequency	Linkage frequency	Linked prior to COVID-19
Australian Immunisation Register (AIR)	Vaccination records	National register that records vaccinations given to everyone in Australia.	Daily	Weekly (soon to be daily)	No
Emergency Department Data Collection (EDDC)	Emergency Department attendances	Patient level data for all public and private emergency department attendances, Only one ICD-10 primary diagnosis code for each attendance.	Daily	Weekly	Yes
Hospital Live Admissions	Hospital admissions	Collects information on admissions and inter-ward transfers in public and private hospitals. Data is rapidly available but not cleaned.	Every 3 hours	Daily	No
Hospital Morbidity Data Collection (HMDC)	Hospital admissions	Reference database for all public and private hospital admissions. Data is cleaned and ICD-10 coded by a data quality and clinical coding team.	Daily	Monthly	Yes
Public and private pathology COVID-19 test results	Pathology	Positive and negative PCR COVID-19 tests, COVID-19 rapid antigen tests and serology results from public and private laboratories.	Twice daily	Daily	Yes
WA Notifiable Infectious Diseases Database (WANIDD)	Infectious diseases	Contains patient information on notifiable infectious diseases	Daily	Daily	Yes
WA Registry of Deaths	Deaths	All deaths are legally required to be reported to the Registry of Deaths. The cause of death and contributing factors are detailed.	Weekly	Fortnightly	Yes
WA Vaccine Safety Surveillance System (WAVSS)	Adverse events following immunisation reports	Reporting service for patients and providers for any adverse events following immunisation.	Daily	Daily	No

Hospital live admissions

Hospital live admissions gives a ‘current snapshot’ of everyone currently in hospital (including emergency departments) in WA and is updated every three hours. This dataset contains limited information on patient demographics, along with data on the hospital and ward. There is a ‘free text’ field with the reason for presentation/admission, however this is not systematically coded and is not reliable for identifying conditions of interest compared to ICD-10 codes in other collections. This collection however has potential for rapid identification of all cause hospital admissions for vaccine safety surveillance.

Hospital Morbidity Data Collection (HMDC)

The HMDC is the quality assured, reference database for hospital admissions. The ICD-10 coding process is undertaken by professional coders and a data quality team. For each hospital admission, an ICD-10 code is assigned for the principal diagnosis, and any number of secondary codes can be allocated. A condition onset flag is allocated for each ICD-10 code to indicate if the event was observed to have occurred during their admission, rather than the person presenting to the hospital with the condition. The HMDC data dictionary can be found through this link.[2]

Pathology data

There are a number of notifiable conditions covered by the *Public Health Act 2016*, including COVID-19, which require mandatory reporting of positive specimens from both public and private laboratories. During the pandemic, data was received multiple times a day to meet reporting demands. However, this frequency will likely reduce to pre-pandemic times. Pathology data contains limited demographic fields along with information on the notifiable disease and is considered highly accurate.

Western Australia registry of deaths

By law, a medical certificate cause of death must be completed within 48 hours of a death in Western Australia and given to the funeral service. The funeral director has 14 days from the funeral service to notify the register [14]. For each death, the

[2] Hospital-Morbidity-Data-Collection-Data-Dictionary-2022.pdf (health.wa.gov.au)

cause of death and contributing factors are written in a free text field and linked to the CVLDR. Due to a long coding delay, ICD-10 coded death data are not available for around two years after the event.

Western Australia Notifiable Infectious Diseases Database (WANIDD)

The WA Notifiable Infectious Diseases database (WANIDD) contains information on notifiable infectious diseases that have been reported by healthcare providers or laboratories, as required by law. This collection is considered the ‘source of truth’ for notifiable infectious diseases, with data transmitted from WANIDD to the National Notifiable Disease Surveillance System. This contains demographic data, along with relevant fields for outbreak management and reporting.

Western Australia Vaccine Safety Surveillance System (WAVSS)

Data from WAVSS contains information from the public or clinician regarding AEFIs, including clinical follow up notes from dedicated WAVSS nurses. This system is used to meet reporting obligations for adverse drug reactions to the national body, the Therapeutic Goods Administration of Australia. As a passive reporting system, this relies on a patient or clinician to report an event, and therefore does not capture every AEFI.

Dataset structure and variables

As the backbone of the CVLDR, individuals within AIR are all assigned a unique identifier. This same identifier is repeated across all data collections in the CVLDR, meaning that all datasets within it are able to be linked together. The collections remain as separate tables, with analysts joining these datasets using the unique identifier to create specific data ‘views’ in order to answer questions of public health significance.

Key demographic variables common across the datasets include age, sex, Aboriginality, and postcode of residence. For a subset of persons who appear in hospitalisation or notifiable disease datasets, information about country of birth, languages spoken, and occupation are also collected. As administrative datasets, there are a large number (>500) of variables with different levels of use for the stated purposes of the CVLDR. The datasets are commonly used to create variables denoting the presence or absence of a condition of interest, such as a notifiable disease, receipt of vaccination or hospitalisation proximal to vaccination or to a positive test.

Discussion

The use of secondary data from large administrative datasets has several known limitations. These data are collected for the purposes of monitoring and reporting on health system activity, rather than epidemiological analysis. Therefore, there are several key variables not available within the linked collection. Potential confounders such as comorbidities may be better assessed using data from General Practice or the Pharmaceutical Benefits Scheme. As these collections are not available in WA for linking, comorbidities cannot be directly controlled for in statistical analyses, hence proxy measures are required.

From a safety monitoring perspective, people in residential aged care or disability care facilities are more likely to be prioritised for some vaccinations (including particular brands of vaccine) and have an increased likelihood of presenting to an emergency department or being admitted to hospital. This is due to either increased medical frailty, or a potentially serious AEFI. Epidemiological analysis in this group may be difficult without the ability to control for existing comorbidities. Potential strategies to address these limitations include restrictions based on age to remove possible aged care residents or using the number of hospitalisation episodes in the 12 months prior to vaccination for select conditions of interest as a proxy measure.

Similarly, the inability to control for differential exposure risk profiles based on occupation (e.g., healthcare or quarantine hotel workers), has implications for the accurate assessment of vaccine effectiveness. Additionally, controlling for the number of COVID-19 polymerase chain reaction (PCR) tests an individual has received, could be used as a proxy measure to identify people in high-risk occupations, who undertake regular testing or screening.

Applications and usage

Applications for vaccine safety

Existing systems to detect safety signals or monitor rare but serious events rely on either a provider or patient to report an AEFI in a timely manner. Data linkage removes the need for third-party reporting of hospital encounters post-vaccination by allowing linkage between the AIR and hospital data collections in WA. This means that cases with potential conditions of interest can be proactively identified if they have contact with the hospital system post vaccination.

The CVLDR has already been used to assist with case finding of adverse events of special interest, including potential thrombosis with thrombocytopenia syndrome cases following COVID-19 vaccination. Linked data from the AIR, inpatient hospitalisation, emergency departments, deaths and WAVSS are consolidated into a single extract which provides aggregated information about vaccine recipients and their post-vaccination contacts with the health system. Clinical experts review this extract to identify suspect AEFIs for further investigation, and helps ensure more complete capture of these cases, which may otherwise be undetected by pre-existing reporting mechanisms. In addition, local protocols to prospectively use linked hospitalisation and emergency department presentation data for signal detection are currently being developed.

Applications for vaccine effectiveness

Along with vaccine safety monitoring, a population-wide linked data collection also has applications for vaccine effectiveness. With the introduction of several new vaccines for COVID-19, there is significant interest in assessing the effectiveness of vaccines against COVID-19 infection, hospitalisation, and death. Additionally, information regarding vaccine effectiveness against variant strains of concern will be important for the planning of the COVID-19 immunisation program in the future.

There are limitations with several of the most common methodologies to assess vaccine effectiveness, including the challenge of enrolling the large sample sizes required, particularly for sub-analyses of specific age groups or certain strains of pathogens [15]. Having access to vaccination, testing, hospitalisation, and mortality data will allow timelier and more precise estimates of vaccine effectiveness. Additionally, future use of data linkage for assessing the effectiveness of the seasonal influenza vaccine is being explored and would allow for data to be analysed earlier in the season.

Future applications

Work is currently underway to develop protocols to use the CVLDR for other immunisations for prospective safety signal detection and assessment of vaccine effectiveness. Methods for the detection of safety signals using emergency and hospitalisation data, similar to those used for post-licensure safety monitoring of vaccines elsewhere [16], are being validated. WA has also joined a global consortium of countries developing methods for rapid cycle analysis called the Global Vaccine Data Network [17]. Assessment of vaccine effectiveness against the SARS-CoV-2 Omicron strain is in progress, and the system is being developed and the system is being developed in preparation for future outbreaks. Another application of the CVLDR is to monitor the impact of 'long COVID' and if prior vaccination reduced a person's risk of ongoing sequalae.

Conclusion

As noted, the applications for this system to assess vaccine safety and effectiveness go beyond the COVID-19 pandemic, with potential access to vaccination, pathology, and morbidity/mortality data for the National Immunisation Program vaccines. The potential inclusion of further data collections which may improve the accuracy of information and the ability to target and better understand subpopulations of interest are also being explored. These include information on healthcare access, comorbidities, and other conditions of interest such as pregnancy, culturally and linguistically diverse populations, disability, residential aged care status and socio-economic status. Building a comprehensive linked data collection will allow WADOH's Immunisation program to maximise the potential of this resource.

Acknowledgements

The authors would like to thank:

- The Linkage Team for linking the data
- The Information and System Performance Directorate at the WADOH, for technical and governance work throughout the development the CVLDR
- State and Commonwealth Data Custodians for granting access to the data
- The Global Vaccine Data Linkage Network for advice on methodologies
- Everyone at the WADOH who supported the original development of the linked data collection for the COVID-19 response

Statement on conflicts of interest

None declared.

Ethics statement

This publication is not subject to ethical approval. It describes the development of infrastructure for Western Australia's response to the COVID-19 pandemic and does not describe the findings of any human research conducted via the infrastructure. Any such research will be discussed in follow-up publications and subject to ethical approval as required.

References

1. Centers for Disease Control and Prevention (CDC). Vaccine Safety Datalink (VSD) 2020 [Available from: https://www.cdc.gov/vaccinesafety/ensuringsafety/monitoring/vsd/index.html.

2. Magnus MC, Oakley L, Gjessing HK, Stephansson O, Engjom HM, Macsali F, et al. Pregnancy and risk of COVID-19: a Norwegian registry-linkage study. BJOG : an international journal of obstetrics and gynaecology. 2022;129(1):101–9. https://doi.org/10.1111/1471-0528.16969

3. Gram MA, Nielsen J, Schelde AB, Nielsen KF, Moustsen-Helms IR, Sørensen AKB, et al. Vaccine effectiveness against SARS-CoV-2 infection, hospitalization, and death when combining a first dose ChAdOx1 vaccine with a subsequent mRNA vaccine in Denmark: A nationwide population-based cohort study. PLoS Med. 2021;18(12):e1003874. https://doi.org/10.1371/journal.pmed.1003874

4. Wolter N, Jassat W, Walaza S, Welch R, Moultrie H, Groome M, et al. Early assessment of the clinical severity of the SARS-CoV-2 omicron variant in South Africa: a data linkage study. The Lancet (British edition). 2022;399(10323):437–46. https://doi.org/10.1016/S0140-6736(22)00017-4

5. Perry M, Gravenor MB, Cottrell S, Bedston S, Roberts R, Williams C, et al. COVID-19 vaccine uptake and effectiveness in adults aged 50 years and older in Wales UK: a 1.2m population data-linkage cohort approach. Human vaccines & immunotherapeutics. 2022;18(1):2031774–. https://doi.org/10.1080/21645515.2022.2031774

6. Carcione D, Blyth CC, Mak DB, Effler PV. User satisfaction with the Western Australian Vaccine Safety

Surveillance (WAVSS) System. Aust N Z J Public Health. 2013;37(3):296. https://doi.org/10.1111/1753-6405.12057

7. Clothier HJ, Crawford NW, Kempe A, Buttery JP. Surveillance of adverse events following immunisation: the model of SAEFVIC, Victoria. Communicable diseases intelligence quarterly report. 2011;35(4):294–8. Available from: https://www1.health.gov.au/internet/main/publishing.nsf/Content/cda-cdi3504d.htm

8. Eitelhuber T, Thackray J, Hodges S, Alan J. Fit for purpose - developing a software platform to support the modern challenges of data linkage in Western Australia. International Journal of Population Data Science. 2018;3(3). https://doi.org/10.23889/ijpds.v3i3.435

9. Eitelhuber T. Data linkage – making the right connections. In: WA Department of Health, editor. Perth, Australia.2016. Available from: https://www.datalinkage-wa.org.au/wp-content/uploads/2019/02/Data-Linkage-Branch-Linkage-Quality.pdf

10. Dalton LG, Meder KN, Beard FH, Dey A, Hull BP, Macartney KK, et al. How accurately does the Australian Immunisation Register identify children overdue for vaccine doses? A national cross-sectional study. Commun Dis Intell (2018). 2022;46. https://doi.org/10.33321/cdi.2022.46.10

11. Hull BP, Lawrence GL, MacIntyre CR, McIntyre PB. Immunisation coverage in Australia corrected for under-reporting to the Australian Childhood Immunisation Register. Aust N Z J Public Health. 2003;27(5):533–8. https://doi.org/10.1111/j.1467-842X.2003.tb00829.x

12. Dalton L, Meder K, Beard F, Dey A, Hull B, McIntyre P, et al. Australian Immunisation Register Data Transfer. Study Stage 2 Final Report. August 2018. Sydney, Australia.: National Centre for Immunisation Research and Surveillance; 2018. https://www.ncirs.org.au/sites/default/files/2018-12/2018%20AIR%20data%20tranfer%20report_FINAL_0.pdf

13. Department of Health. Mandatory reporting of National Immunisation Program vaccines to the Australian Immunisation Register began on 1 July 2021 2021 [Available from: https://www.health.gov.au/news/mandatory-reporting-of-national-immunisation-program-vaccines-to-the-australian-immunisation-register-began-on-1-july-2021.

14. Government of Western Australia. Apply for a death certificate 2022 [Available from: https://www.wa.gov.au/service/justice/civil-law/apply-death-certificate.

15. World Health Organization (WHO). Evaluation of COVID-19 vaccine effectiveness. Interim Guidance. 12 March 2021. Geneva, Switzerland.; 2021. Available from: https://apps.who.int/iris/handle/10665/340301

16. Baggs J, Gee J, Lewis E, Fowler G, Benson P, Lieu T, et al. The Vaccine Safety Datalink: A Model for Monitoring Immunization Safety. Pediatrics. 2011;127(Supplement_1):S45–S53. https://doi.org/10.1542/peds.2010-1722H

17. Global Vaccine Data Network. Global Vaccine Data Network - About Us 2022 [Available from: https://www.globalvaccinedatanetwork.org/aboutus.

Abbreviations

AEFI:	Adverse Event Following Immunisation
AIR:	Australian Immunisation Register
csv:	comma-separated value
CVLDR:	COVID-19 Vaccination Linked Data Repository
EDDC:	Emergency Department Data Collection
HMDC:	Hospital Morbidity Data Collection
PCR:	Polymerase chain reaction
PRC:	Pandemic Response Collection
VE:	Vaccine Effectiveness
WA:	Western Australia
WADOH:	Western Australian Department of Health
WANIDD:	Western Australia Notifiable Infectious Diseases Database
WAVSS:	Western Australian Vaccine Safety and Surveillance