Australian Government

Australian Institute of Health and Welfare

COVID-19 Register: linkage results

Web report | Last updated: **07 Mar 2024** | Topic: COVID-19

About

The COVID-19 Register is a linked data asset that enables evidence-based research, and public health and health system planning for current and future pandemics. This linkage project also facilitates selected information to be fed back and incorporated into the local notifiable disease systems of states and territories to improve data completeness and utility.

The first version of the linked data was first made available to approved researchers in December 2022. The linked data has since been updated to Version 2 and Version 2.5.

The purpose of this report is to briefly describe the linkage method and process, and compare high level linkage results for previous and current versions of the COVID-19 Register. This report will be a resource for future researchers wanting to use the data.

Cat. no: PHE 328

Findings from this report:

- The AIHW has expanded the coverage of the COVID-19 Register to include over 7 million linked COVID-19 cases

- Linkage rates for Version 2.5 were generally similar to Version 2, with over 90% of records linked

- Linkage rates can differ slightly by jurisdictions and population groups

Data and methods

Ethics approvals

The project has obtained ethics approval from the AIHW Ethics Committee, and additional approval from the Human Research Ethics Committee of Northern Territory Department of Health and Menzies School of Health Research, and the New South Wales Population and Health Services Research Ethics Committee (NSW PHSREC). A National Mutual Acceptance Scheme led by NSW PHSREC is in place for the Australian Capital Territory, South Australia, Tasmania, and Victoria. A data disclosure agreement with Queensland was established for the purpose of this project.

In addition to the ethics approvals outlined above, approval has also been received from the data custodian of each state/territory or national dataset.

How were the data linked?

As an Accredited Data Service Provider, the AIHW is accredited to provide complex integration, de-identification and secure access to linked data.
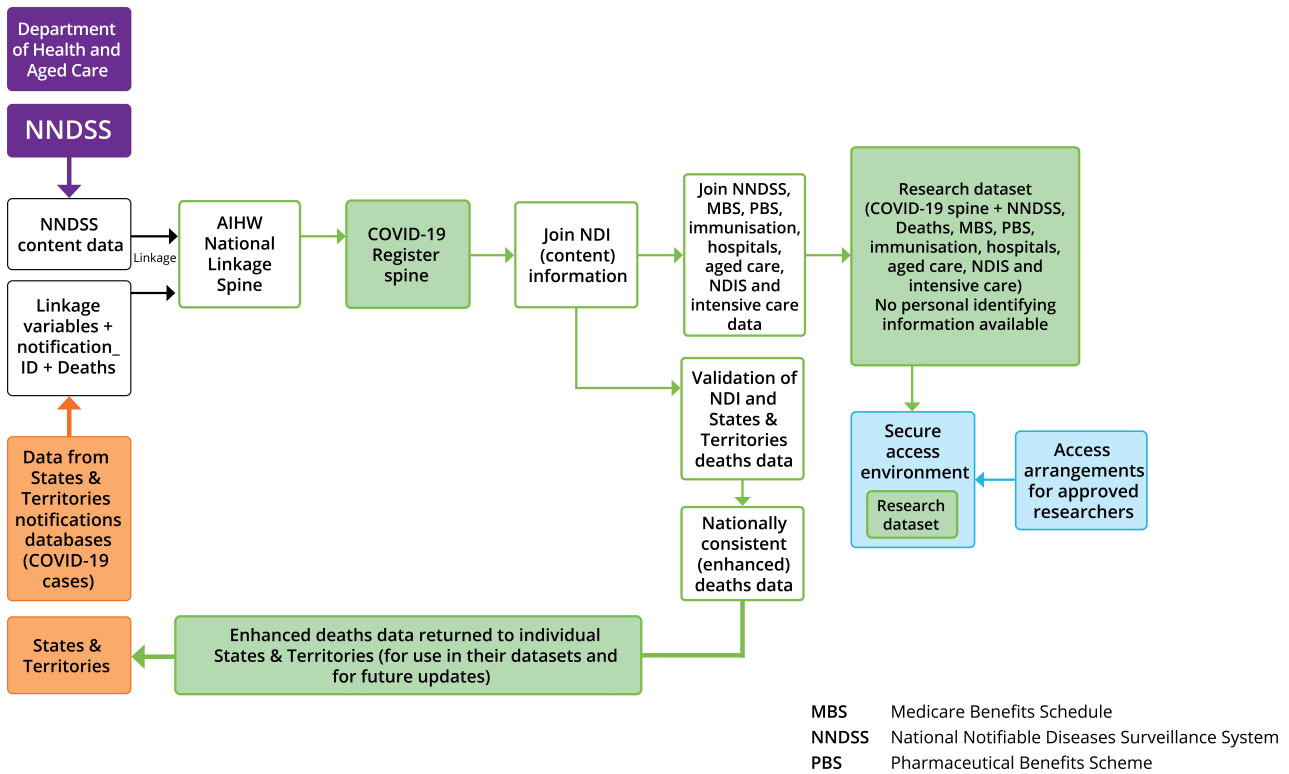
COVID-19 case linkage variables (names, addresses, dates of birth and sex) provided by jurisdictions to the AIHW were probabilistically linked to the AIHW National Linkage Spine (NLS). The AIHW NLS combines linkage variables from Medicare Consumer Directory (MCD), National Death Index (NDI), Australian Immunisation Register (AIR) and uniquely covers almost all the population of Australia. Probabilistic record linkage is a data linkage method that makes an explicit use of probabilities to determine whether a pair of records is a match for the same person, or not. Records are matched by name, sex, address and date of birth. The resulting COVID-19 Register does not, however, contain any identifying information.

Analytical information on COVID-19 cases from states and territories and the Commonwealth Department of Health and Aged Care National Notifiable Disease Surveillance System (NNDSS) were combined with information from the NDI, Medical Benefits Schedule (MBS), Pharmaceutical Benefits Scheme (PBS), the National Hospitals Morbidity Database (NHMD), the National Non-Admitted Patient Emergency Department Care Database (NNAPEDCD), the National Aged Care Data Clearinghouse (NACDC), the Australian and New Zealand Intensive Care Society (ANZICS), the AIR and the National Disability Insurance Scheme (NDIS) to create a de-identified linked research data set. Figure 1 outlines the linkage processes for the current version of the project (Version 2.5).

After both the initial and re-linkages, date of death and cause of death information from the NDI is released to the states and territories that provide the original notifiable disease data, for incorporation into their local notifiable disease systems. The aim of this is to improve NNDSS data completeness and utility, in a nationally consistent way, and add to the research potential of both the state and territory collections and the NNDSS.

The AIHW data linkage protocols are based on the Five Safes framework which reinforce management of the privacy and confidentiality of data. These protocols prescribe strict separation of identifiers and analytical data. This means AIHW linkage staff do not have access to the personal identifiers and analytical data at the same time for the duration of the project.

Figure 1: COVID-19 linked data flow

Department of Health and Aged Care

NNDSS

NNDSS content data

Linkage variables + notification_ID + Deaths

Linkage

Data from States & Territories notifications databases (COVID-19 cases)

States & Territories

AIHW National Linkage Spine

COVID-19 Register spine

Join NDI (content) information

Join NNDSS, MBS, PBS, immunisation, hospitals, aged care, NDIS and intensive care data

Validation of NDI and States & Territories deaths data

Nationally consistent (enhanced) deaths data

Enhanced deaths data returned to individual States & Territories (for use in their datasets and for future updates)

Research dataset (COVID-19 spine + NNDSS, Deaths, MBS, PBS, immunisation, hospitals, aged care, NDIS and intensive care) No personal identifying information available

Secure access environment

Research dataset

Access arrangements for approved researchers

MBS	Medicare Benefits Schedule
NNDSS	National Notifiable Diseases Surveillance System
PBS	Pharmaceutical Benefits Scheme

A longitudinal resource for COVID-19 cases

The COVID-19 Register will be updated to include case data up to December 2022 for all jurisdictions. This date corresponds with the relaxation in testing and reporting requirements, which limit the completeness of data after this time. For deaths data, Australian Bureau of Statistics coded cause of death information will be incorporated as it becomes available. Updating the content provides a growing longitudinal resource for COVID-19 cases and allows research into the patients' health journey over time. Providing data back to the states and territories will enhance data completeness of their notifiable disease systems.

Linkage findings

Scope of the data

Version 1 of the COVID-19 Register had around 250,000 records linked to a range of administrative data sets. Version 2 had an expanded data coverage to more than 6 million linked records with more recent case data (New South Wales), more jurisdictions (Victoria and Queensland) and more data sets. Version 2.5 includes case data for Tasmania, Northern Territory, Victoria and the Australian Capital Territory to 31 December 2022 and updated hospitals data for the Australian Capital Territory, Victoria, New South Wales, Queensland and Tasmania (to June 2022). Table 1 provides an overview of the data sets and coverage between the versions.

Please refer to the data variables list for the temporal scope of each of the datasets and how it differs between versions.

Table 1: List of data sets included between versions

Data set	Version 1 (released in December 2022) *Number of linked records = 250,821*	Version 2 (released in November 2023) *Number of linked records = 6,415,740*	Version 2.5 (released in February 2024) *Number of linked records = 7,256,727*
State/territory notifiable disease data on COVID-19 cases	ACT NSW NT SA Tas	ACT NSW (updated data) NT SA Tas Vic (new) Qld (new)	ACT (12/03/20–31/12/22) NSW (25/01/20–30/09/22) NT (21/02/20–31/12/22) SA (30/01/20–11/02/22) Tas (30/03/20–31/12/22) Vic (25/01/20–31/12/22) Qld (28/01/20–20/08/22)
Medicare Consumer Directory (MCD)	Yes Whole of population	Yes Whole of population	Yes Whole of population
National Death Index (NDI)	Yes Whole of population	Yes Whole of population	Yes Whole of population
Medicare Benefits Schedule (MBS)	Yes Cases only	Yes Whole of population	Yes Whole of population
Pharmaceutical Benefits Scheme (PBS, including Repatriation Schedule of Pharmaceutical Benefits (RPBS) information)	Yes Cases only	Yes Whole of population	Yes Whole of population
Australian Immunisation Register	Yes Whole of population	Not available	Yes Whole of population
National Notifiable Disease Surveillance System (NNDSS)	Yes Cases only	Yes Cases only	Yes Cases only
National Hospitals Morbidity Database (NHMD)	Yes Cases only	Yes Whole of population	Yes Whole of population
National Non-Admitted Patient Emergency Department Care Database (NNAPEDCD)	Yes Cases only	Yes Whole of population	Yes Whole of population
National Aged Care Data Clearinghouse (NACDC)	Yes Cases only	Yes Whole of population	Yes Whole of population
National Disability Insurance Scheme (NDIS)	Not available	Not available	Yes Whole of population

| Australian New Zealand Intensive Care Survey (ANZICS) Adult Patient Database (APD)	Not available	Not available	Yes Whole of population
Australian and New Zealand Paediatric Intensive Care Registry (ANZPICR)	Not available	Not available	Yes Whole of population

Linkage rates by jurisdiction

Generally, linkage results depend on the accuracy and completeness of the linkage variables provided to the AIHW: more accurate and complete data result in better linkage rates. For more information on how the data are linked, please refer to the above section on Data and methods.

Figure 2 shows the number of records that were linked and those that were unable to be linked by state and territory. For most jurisdictions, linkage rates have generally remained the same or improved slightly, where over 90% of records supplied for the project were linked in Versions 1, 2 and 2.5. The exception is Tasmania where the proportion of linked cases fell from 99% in Version 2 to 89% in Version 2.5. This is due to the notable increase in cases from Version 2 (243) to Version 2.5 (282,277), and the high proportion of cases with missing address information, for example, 63% of cases were missing the 'city' variable, and 74% were missing the 'street1' variable.

There were notable increases in the number of records supplied from the Australian Capital Territory and Northern Territory between Version 2 and Version 2.5. The linkage rate improved slightly, from 96% to 98% for the Australian Capital Territory and from 93% to 97% for the Northern Territory. New data supplies for South Australia, Queensland and New South Wales will be reflected in the next version of the COVID-19 Register, anticipated in April 2024.

Figure 2: Number of records and percentage linked by jurisdictions across versions

This bar chart shows the linkage rates for most jurisdictions have remained the same or improved slightly across versions of the COVID-19 Register.
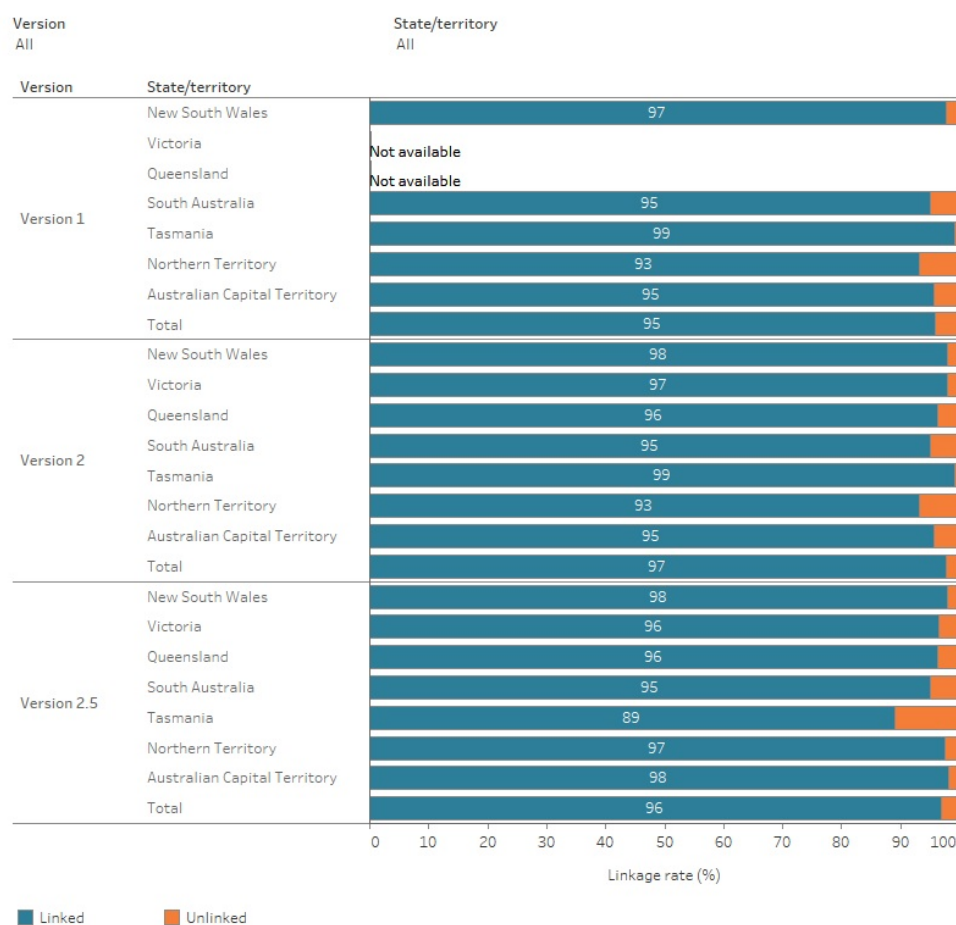
Note: Results for Version 1 are based on participating states and territories and will not be directly comparable to the figures in the previous web report 'Establishing a COVID-19 linked dataset', which includes Victoria (released 16 December 2022).
Source: COVID-19 Register
https://www.aihw.gov.au

Linkage rates by population groups

Table 2 describes the linkage rates by age group and sex/gender. Linkage rates can differ by population groups, and it is important to consider this when conducting analysis on linked data. Table 2 shows that the linkage rate largely improved for Version 2 compared to Version 1, where the linkage rate for all groups was well over 90%, except the 'Other' sex/gender category. This has remained similar for Version 2.5. Sex is one of the key variables used to link records, therefore, where sex is not reported consistently, or as neither male nor female ('Other' in Table 2 below) linkage rates are lower. The linkage rate for 'Other' considerably improved from 3% in Version 1 to about 77% in Version 2 and Version 2.5, though the linkage rate remains lower than males or females. There were no other large differences observed in linkage rates across the age groups.

Table 2: Number of records and percentage linked by population groups across versions

	Version 1[1]	Version 2	Version 2.5
Sex/gender[2]			
Male	125,673 (96.4%)	3,020,677 (97.7%)	3,403,063 (96.8)
Female	125,075 (97.2%)	3,382,173 (97.8%)	3,836,445 (97.0)
Other[3]	73 (3.0%)	13,765 (77.3%)	18,433 (77.5)
Age group[4]			
0-15	47,241 (96.6%)	1,141,652 (97.2%)	1,281,183 (95.8)
16-29	73,074 (95.1%)	1,463,851 (97.2%)	1,638,564 (96.0)
30-49	79,326 (95.9%)	2,104,378 (98.3%)	2,373,946 (97.4)
50-69	39,433 (96.9%)	1,253,801 (98.8%)	1,430,191 (97.9)
70+	11,747 (95.9%)	452,888 (94.7%)	534,007 (96.8)

Additional data on linkage rates by population groups are available in the supplementary tables.

1. Results for Version 1 (released on 16 December 2022) are based on those participating states and territories as detailed in Figure 2 and will not be directly comparable to the figures in the previously released web report 'Establishing a COVID-19 linked dataset' which also includes Victoria.
2. As reported by the state and territory.
3. Other includes records where sex or gender is not reported, or sex is reported as neither male nor female.
4. Age group is based on age as at 31 December 2022. Records with missing information on birth date are excluded. Person IDs with more than one year of birth and/or sex were restricted to the most recent notification date (only small number of records were affected). Where the notification dates were equal, a random record was used.

Future developments

The next steps of this project are to improve the coverage of COVID-19 cases, which includes linking case data up to December 2022 for all jurisdictions.

The COVID-19 Register has been built to ensure interoperability with existing national data assets, such as the AIHW's National Integrated Health Services Information (NIHSI) and the Australian Bureau of Statistics' Person Level Integrated Data Asset (PLIDA). During 2024, the COVID-19 case data from the COVID-19 Register will be integrated with the NIHSI, which is being re-branded as the National Health Data Hub (NHDH). The NHDH aims to re-use current data infrastructure to support and streamline broader data integration, bring AIHW data holdings together using a modular approach as needed and as approved per project, and most importantly reduce duplication of effort. This integration will assist with ensuring the on-going use of the COVID-19 case data at the end of the current funding period (December 2024).

The AIHW and Australian Bureau of Statistics are also working on the Australian National Data Integration Infrastructure (ANDII). This is a national linkage and integration infrastructure built to streamline and reduce the time required to build integrated datasets. More information can be found at: National Disability Data Asset, Australian Bureau of Statistics (abs.gov.au).

Such national platforms can help facilitate the sharing of data between government agencies as well as facilitate the integration of data from various sources and streamline efforts in data linkage.

Abbreviations

AIR	Australian Immunisation Register
ANDII	Australian National Data Integration Infrastructure
ANZICS	Australian and New Zealand Intensive Care Society
ANZICS APD	Australian New Zealand Intensive Care Survey Adult Patient Database
ANZPICR	Australian and New Zealand Paediatric Intensive Care Registry
MBS	Medicare Benefits Schedule
MCD	Medicare Consumer Directory
NACDC	National Aged Care Data Clearinghouse
NDI	National Death Index
NDIS	National Disability Insurance Scheme
NHDH	National Health Data Hub
NHMD	National Hospitals Morbidity Database
NIHSI	National Integrated Health Services Information
NNAPEDCD	National Non-Admitted Patient Emergency Department Care Database
NNDSS	National Notifiable Disease Surveillance System
PBS	Pharmaceutical Benefits Scheme
PLIDA	Person Level Integrated Data Asset
RPBS	Repatriation Schedule of Pharmaceutical Benefits

Archived content

Data