# COVID-19 linked data set; Quality Statement

## Exported from METEOR (AIHW's Metadata Online Registry)

# COVID-19 linked data set; Quality Statement

## Identifying and definitional attributes

**Metadata item type:** Data Quality Statement

**METEOR identifier:** 768139

**Registration status:** AIHW Data Quality Statements, Superseded 17/08/2023

## Data quality

**Data quality statement summary:**

Summary

- Participation and contribution to the COVID-19 linked data set by jurisdictions is voluntary. The first version of this data set includes COVID-19 case notification data from New South Wales (NSW), South Australia (SA), Tasmania (Tas), Northern Territory (NT), and the Australian Capital Territory (ACT) with remaining states to be provided in future updates.
- While the Australian Institute of Health and Welfare (AIHW) continues to explore avenues to secure approvals for data sharing with all jurisdictions, analysts should note limitations with the coverage of hospitals data in this first version of the data set, in particular:
    - Hospitalisation data from Western Australia and the Northern Territory are not included
    - Records from private hospitals are not captured in this version but will be considered in future
    - The current coverage period for hospitals data is 01/07/2014 – 30/6/2021 with 2021-22 data to be included in the next update.
- The data set includes service-based data sources (such as Pharmaceutical Benefits Scheme (PBS), Medical Benefits Schedule (MBS) and the Australian Immunisation Register (AIR)). It consists of records of services provided to people who are usual residents of Australia. It may also capture some people who live in Australia but are not eligible for Medicare (e.g., international students, visitors to Australia from countries with reciprocal healthcare agreements). In addition, everyone in Australia is eligible for a free COVID-19 vaccination and as such immunisation records will capture individuals who are not usual residents of Australia. Both under coverage and over coverage of different cohorts of interest within the Australian resident population need to be considered in the analysis and interpretation of the data set.
- The data set has received ethics approval for whole of population coverage for the Medicare Consumer Directory (MCD), Pharmaceutical Benefits Scheme (PBS) and Medical Benefits Schedule (MBS) data sets. However, due to the size and complexity of the data, only PBS and MBS data for COVID-19 cases will be available in this first version, whilst MCD will cover the whole population.
- Each state and territory may have unique testing and reporting requirements based on jurisdictional public health orders. For jurisdictions where rapid antigen test (RAT) results are not included in case notifications linked to the COVID-19 linked data set, there will be an underreporting of COVID-19 positive cases.
- RATs performed in a home setting are open to user error and the obligation to record positive results rests with patients, who may not be aware of this duty, or of reporting mechanisms. This may also result in underreporting of COVID-19 cases.
- Due to underreporting of RAT and/or PCR results, analysts using this data set should be aware that there might be individuals who may have never been a notified case in state and territory notifiable diseases. However, they may

have died from COVID-19. These individuals will enter the cohort through the National Deaths Index (NDI) data source where emergency codes relating to COVID-19 were used to code cause of death.

Description

The COVID-19 linked data set brings together COVID-19 cases from states and territories and the Commonwealth Department of Health National Notifiable Disease Surveillance System (NNDSS) combined with national health administrative datasets including the:

- Medicare Consumer Directory (MCD)
- National Deaths Index (NDI)
- Medical Benefits Schedule (MBS)
- Pharmaceutical Benefits Scheme (PBS, including Repatriation Schedule of Pharmaceutical Benefits (RPBS) information)
- National Hospitals Morbidity Database (NHMD)
- National Non-Admitted Patient Emergency Department Care Database (NNAPEDCD)
- National Public Hospitals Establishment Database (NPHED)
- National Aged Care Data Clearinghouse (NACDC) and
- Australian Immunisation Register (AIR).

The COVID-19 linked data set aims to provide greater insights into the longer-term impact of COVID-19 on the health of the Australian population and the health system. Research outcomes are anticipated to inform health service planning, monitoring and evaluation purposes and policy development. The data set can be accessed by approved analysts through a secure remote environment.

**Institutional environment:**

The AIHW develops, maintains and manages the use of the COVID-19 linked data set. The Australian Institute of Health and Welfare (AIHW) is an independent corporate Commonwealth entity under the Australian Institute of Health and Welfare Act 1987 (AIHW Act), governed by a management Board and accountable to the Australian Parliament through the Health portfolio.

The AIHW is a nationally recognised information management agency. Its purpose is to create authoritative and accessible information and statistics that inform decisions and improve the health and welfare of all Australians.

Compliance with the confidentiality requirements in the AIHW Act, the Privacy Principles in the Privacy Act 1988 (Cth) and AIHW's data governance arrangements ensures that the AIHW is well positioned to release information for public benefit while protecting the identity of individuals and organisations.

For further information see the AIHW website https://www.aihw.gov.au/about-us, which includes details about the AIHW's governance (https://www.aihw.gov.au/about-us/our-governance and our role and strategic goals (https://www.aihw.gov.au/about-us/what-we-do)

The COVID-19 linked data set is funded by the Medical Research Future Fund (MRFF). AIHW engages in quarterly consultations with members of the COVID-19 Data Advisory Group represented by members from state/territory health departments, the Commonwealth Department of Health and the National Centre for Immunisation Research and Surveillance. The role of the COVID-19 Register Advisory Group (the Advisory Group) is to provide a wide range of expert advice to AIHW regarding the COVID-19 linked data set.

**Timeliness:**

The first version of the COVID-19 linked data set was completed in December 2022. In the first stage of the project, only government or government-funded researchers will be eligible to apply for access. All other researchers will have access to the data in future stages of the project once all relevant ethics and data custodian approvals for access and use arrangements have been obtained.

Timing of updates to the COVID-19 linked data set will be subject to timely provision of case notification data to the AIHW from state and territory governments, and timely access to the content data via the administrative datasets(s), and subject to agreement from data custodians.

The project aims to re-link information periodically to identify additional deaths, and to update data where available. ABS coded cause of death information will be incorporated as it becomes available.

Population and coverage periods of the data sources that make up the COVID-19 linked data set are listed in Table 1.

Table 1. Population and coverage period for Version 1 of the COVID-19 linked data set

| Data Set | State/Territory | Coverage period |
|---|---|---|
| Medicare Consumer Directory (MCD) | National (whole of population) | Extracted up to 30/6/2022 |
| Pharmaceutical Benefits Scheme (PBS) | National (cases only) | 01/01/2015 – 30/6/2022 |
| Medicare Benefits Schedule (MBS) | National (cases only) | 01/01/2015 – 30/6/2022 |
| National Notifiable Diseases Surveillance System (NNDSS) | NSW, SA, Tas, NT, ACT (cases only) | 01/01/2020 – 30/6/2022 |
| National Death Index (NDI) | National (whole of population) | 01/01/2020 – 30/6/2022 |
| Australian Immunisation Register (AIR) | National (whole of population) | 01/01/2020 – 30/6/2022 |
| National Hospitals Morbidity Database (NHMD)-public hospitals only | NSW, Vic. Qld, SA, Tas. ACT (cases only) | 01/07/2014 – 30/6/2021 |
| National Non-Admitted Patient Emergency Department Care Database (NNAPEDCD) | NSW, Vic. Qld, SA, Tas. ACT (cases only) | 01/07/2014 – 30/6/2021 |
| National Public Hospitals Establishment Database (NPHED) | Hospitals that participate in the hospitals establishment database, i.e.<br><br>NSW, Vic, Qld, SA, Tas, ACT (cases only) | 01/07/2014 – 30/6/2021 |
| Aged Care Funding Instrument (ACFI) | All jurisdictions (cases only) | 01/01/2015 – 30/6/2022 |
| National Screening and Assessment Form (NSAF) | | 01/01/2015 – 30/6/2020 |
| Permanent and respite residential aged care (RAC) | | 01/01/2015 – 30/6/2022 |
| Commonwealth Home Support Program (CHSP) | | 01/01/2015 – 30/6/2020 |
| Home Care Packages (HCP) | | 01/01/2015 – 31/3/2022 |

| | | |
|---|---|---|
| Transition Care Program (TCP) | | 01/01/2015 – 30/6/2022 |
| Short-Term Restorative Care (STRC) | | 01/01/2015 – 30/6/2022 |
| State and territory notifiable disease databases | NSW: Notifiable Conditions Information Management System (NCIMS) | 25/01/2020 – 24/11/2021 |
| | SA: Notifiable Infectious Disease Database (NIDD) | 30/01/2020 – 11/02/2022 |
| | Tasmania: Tasmanian Notifiable Disease Database (TNDD) | 30/03/2020 – 30/11/2021 |
| | ACT: ACT Notifiable Diseases Register | 12/03/2020 – 23/03/2022 |
| | NT: NT Notifiable Diseases Surveillance System (NTNDS) | 21/02/2020 – 25/03/2022 |

**Accessibility:**

In the first stage of the project, only government or government-funded researchers will be eligible to apply for access. All other researchers will have access to the data in future stages of the project once all relevant ethics and data custodian approvals for access and use arrangements have been obtained. It should be noted that the current ethics approval only allows the data to be accessed by researchers located in Australia. If there is a need to allow access to overseas researchers in future, relevant ethics approvals will need to be obtained.

The current version of a data variable list can be viewed on the COVID-19 linked data website.

Where data is not publicly available, you may request data by following the steps outlined on the Data on request page on the AIHW website in the first instance. Alternatively, the COVID-19 project team can be contacted on covid19register@aihw.gov.au. Any data request will need to be approved by the relevant data custodians.

All AIHW-authored reports and publication products derived from the use of the COVID-19 linked data set, satisfying output requirements and approval processes will be published and accessible from the AIHW website (www.aihw.gov.au). Publications derived from external researchers will be referenced on the AIHW website once work is published in the public domain.

**Interpretability:**

Information on the COVID-19 linked data set which can help provide insight into the data include data variable lists updated at regular intervals, a tips and tricks document, fact sheets, and web reports with summary outputs derived from the linked data set. These will be made available publicly on the COVID-19 linked data website in December 2022 and thereafter. Where available, metadata for each underlying data source will be published in the AIHW's online metadata repository – METeOR which can be accessed on the AIHW website at METeOR home (aihw.gov.au).

National notification data on COVID-19 confirmed cases is collated in the National Notifiable Diseases Surveillance System (NNDSS) based on notifications made to state and territory health authorities under the provisions of their relevant public health legislation. NNDSS case notifications and variables of interest are available for the 5 participating jurisdictions (NSW, SA, Tas, NT and ACT) in the first version of the linked data set. Future versions will include the remaining jurisdictions. Metadata relating to State/Territory and NNDSS variables used in the COVID-19 linked data set will be available on Meteor (aihw.gov.au).

The National Death Index (NDI) is maintained by AIHW and its data quality statement can be found on the AIHW website at National Death Index (NDI), Data Quality Statement (aihw.gov.au).

The data quality statements underpinning the AIHW National Mortality Database can be found in the following Australian Bureau of Statistics (ABS) publications:

Deaths, Australia methodology, 2020 | Australian Bureau of Statistics (abs.gov.au)

Causes of Death, Australia methodology, 2020 | Australian Bureau of Statistics (abs.gov.au)

Data quality statement (DQS) are not specifically available, however, the metadata on Medicare Benefits Schedule (MBS) data collection and Pharmaceutical Benefits Scheme (PBS) can be found on the AIHW website.

Data quality statements for the AIHW National Aged Care Data Clearinghouse and Aged Care Funding Instrument can be found on METeOR.

The data quality statements for hospitals data can be found in

About the data - Australian Institute of Health and Welfare (aihw.gov.au)

**Relevance:**

The COVID-19 linked data set was created in response for current and ongoing monitoring of the health outcomes and health system needs of people who have had a COVID-19 diagnosis.

There is potential for the linked data set to be used for research under the following approved themes:

- Epidemiological and statistical research
- Service use and medication dispensing and patient journeys
- Identifying groups or cohorts of interest
- Monitoring, evaluation and data quality improvement

The COVID-19 linked data set is national, person-based linked data set, comprising at its core a register of people with a positive diagnosis of COVID-19 recorded in the first three years (pending data supply) after the initial case was recorded in Australia.

The COVID-19 case cohort is derived from state and territory notifiable diseases databases and the NDI. It includes all people tested (via PCR and RAT) who had returned one or more positive results for COVID-19 (SARS-CoV-2 positive) at the time of data extract. That is, where the Commonwealth DISEASE_CODE = 081 (COVID-19) in state and territory notifiable diseases data, or where cause of death on the NDI identifies a COVID-19 diagnosis.

The following groups are excluded from the COVID-19 case cohort and will instead be used to define a comparison group of non-cases. That is, where the Commonwealth DISEASE_CODE is not 081 (COVID-19) in state and territory notifiable diseases data. This comparison group is derived from the Medicare Consumer Directory (MCD) and incudes all people alive at 1/1/2020 who:

- did not return a positive test result for COVID-19 at the time of data extract (these people are not included in the state and territory notifiable diseases data and therefore are excluded)
- had not been tested at the time of data extract (these people are not included in the state and territory notifiable diseases data and therefore excluded)
- are not in the notifiable diseases data cohort, who died and did not have cause of death = U07.1 “COVID-19 virus identified - used when COVID-19 is confirmed by laboratory testing” in ABS-coded cause of death information.

The reference period of each data source and list of participating jurisdictions in the first version of the COVID-19 linked data set is provided in Table 1.

Data on a person’s usual residence varies across data sources in the COVID-19 linked data set, with the minimum geography available at the postcode level or Statistical Area Level 2 (based on the ABS Australian Statistical Geography Standard). SA2 level information is required to analyse data by LGA, and for the derivation of socioeconomic, remoteness and other area classifications commonly used to make comparisons by region (such as SA3 and SA4).

Geographical detail at the postcode and SA2 level instead of individual addresses, are provided in the following data sources:

- SA2: MCD, NDI, NHMD, NNAPEDCD, AIR and NACDC
- Postcode: States/territories case notification, NHMD, NNAPEDCD, and NACDC

Public hospital admitted patient episode data are drawn from the National Hospital Morbidity Database (NHMD). Some admitted patients may not be enrolled in or are eligible for Medicare but will still be included in the NHMD such as international students or some overseas visitors who were admitted to public hospitals. For example, overseas visitors from New Zealand, Ireland, the United Kingdom, the Netherlands, Sweden, Finland Norway, Italy, Malta, Belgium and Slovenia may receive public hospital care because Australia has Reciprocal Health Care Agreements with these countries. Over coverage in these cases may occur due to a lack of information or when the individuals leave Australia and no longer considered as usual residents. This may mean that individuals may continue to be counted in the analysis after they these individuals are no longer resident in Australia unless, methods are applied to adjust for this over coverage. As such under-coverage or over coverage of different groups within the Australian resident population need to be considered in the analysis and interpretation of data. Hospital data are reported based on state of service and not state of usual residence.

Data on admitted patient private hospitals is not available in the first version of the COVID-19 linked data set and as such, is unrepresented in any potential analyses and outputs. Data custodian approval will be sought to identify organisations (hospitals) in research outputs, and no outputs identifying hospitals will be released without such approval.

Data on Indigenous status are available in the NNDSS, NHMD, NNAPEDCD, NACDC, and AIR data sources.

**Accuracy:**

Data included in the COVID-19 linked data set is sourced from AIHW data holdings for MBS, PBS, hospitals, residential aged care, and national deaths data. The data collection and cleaning processes varies across these collections, and the subsequent quality of the linked data set will be subject to the quality of the data held in these source collections. It is important for researchers to recognise that statistical outputs for analysis are generally not the primary reason for the collection of these administrative data.

Data linkage was undertaken using probabilistic linkage, involving creation of record pairs by combining records from one data set with records from another data set based on similarities in characteristics such as last name, first name(s), date of birth, sex, and address of residence. Matches are evaluated based on the level of similarities between the characteristics. A higher level of similarities suggests that a given record pair is more likely to be the same person and treated as a true link.

Quality of linkage depends on the coverage and quality of identifiers available for each collection, and consistency with information held in the integrating spine (i.e.,

MCD, NDI and AIR data were first linked to create the linkage spine).

Link accuracy for the COVID-19 linked data set was a high priority. The addition of AIR data to the MCD-NDI linkage spine increased the rate of successful linked individuals across all jurisdictions. Unlinked records were identified and retained in the data set. This allows analysts the opportunity to conduct sensitivity analyses using characteristics of the unlinked cohort, if required.

The AIHW conducted a program of testing and validation to ensure the integrity and quality of the data set. These checks include:

- completeness of records i.e., identifying missing values or duplicates
- alignment of broad aggregates between the linked data set and the source datasets
- comparability of outputs against published sources of information

Data that are found to be missing, duplicated or are potential outliers will be identified and provided to analysts in a user guide document to assist in their analytical processes.

Jurisdiction-specific accuracy issues include the possible inconsistencies in the way demographic variables like ‘sex’ is coded at data collection i.e., whether this intended as sex at birth or gender and the possibility that this may not be understood by respondents.

Analysts who wish to use indigenous status data for statistical reporting purposes should note the small number of events in selected jurisdictions.

Table 2 shows the linkage rates for the states and territories (NSW, SA, Tas, ACT and NT) datasets that were linked to the MCD-NDI-AIR spine.

Table 2. Number of records and percentage linked by jurisdiction

| State and territory | New South Wales | South Australia | Tasmania | Australian Capital Territory | Northern Territory |
|---|---|---|---|---|---|
| Scope: Start date | 25/01/2020 | 30/01/2020 | 30/03/2020 | 12/03/2020 | 21/02/2020 |
| Scope: End date | 24/11/2021 | 11/02/2022 | 30/11/2021 | 23/03/2022 | 25/03/2022 |
| Number of "linkable" Individuals | 80,357 | 122,654 | 243 | 46,105 | 12,057 |
| Number of linked individuals using MCD-NDI spine alone | 75,638 | 111,856 | 236 | 41,130 | 10,699 |
| Number of linked individuals using MCD-NDI-AIR spine | 78,568 | 116,668 | 241 | 44,114 | 11,230 |
| % Linked individuals | 97.8% | 95.1% | 99.2% | 95.7% | 93.1% |
| % "linked with confidence" individuals | 96.6% | 91.7% | 95.1% | 94.5% | 87.3% |

**Coherence:**

The linked data set is anticipated to be updated quarterly, where data is available. Differences in scope, reference periods and variables (or data sources) between each version of the linked data set will be captured in updated versions of the data quality statement. This should be considered when comparing outputs with different reference periods of the linked data set.

Researchers should note that the hospitals data in the linked data set is derived from existing linkages used in the National Integrated Health Services Information Analysis Asset (NIHSI AA) project. As such, the scope of hospitals data (NHMD, NNAPEDCD and NPHED) may be different to unlinked hospitals data.

Personal project numbers are assigned for each unique individuals in this linked data set and as such, they remain specific and relevant only to this cohort. Researchers must not attempt to link the COVID-19 linked data set to other sources of data.

Demographic data, such as sex, usual residence and/or indigenous status may be captured differently across source data collections. States/territories providing COVID-19 case notification data have different reporting conditions in capturing information on COVID-19 diagnosis (such as the capturing of RAT and/or PCR results) and COVID-19 hospitalisations or death. Work is underway to capture and understand the nuances in the reporting of these variables to be able to advise analysts when comparing data or statistics to other data collections. It is also important to note the different coverage periods between each jurisdiction in the first version of the linked data as this can impact comparability both within the linked data set and between other sources of data collection.

## Data products

**Implementation start date:** 16/12/2022

## Source and reference attributes

**Submitting organisation:** Australian Institute of Health and Welfare

## Relational attributes

**Related metadata references:**

Has been superseded by COVID-19 Register; Quality Statement
AIHW Data Quality Statements, Superseded 07/03/2024

See also National Integrated Health Service Information Analysis Asset (NIHSI AA) version 1.0
AIHW Data Quality Statements, Superseded 21/03/2024