Australian Government
Australian Institute of Health and Welfare

Establishing a COVID-19 linked data set

Web report | Last updated: **16 Dec 2022** | Topic: COVID-19

About

This report has been archived. Content previously included in this report can now be found at *COVID-19 register and linked data set* and *COVID-19 linked data set: Linkage results*.

This is an initial summary report on the project 'Establishing a COVID-19 linked data set'. The purpose of the report is to describe the aims, benefits, methods and some high level first results of the linkage. The report will be a resource for future researchers wanting to use the data. Future updates of this report will include more detailed analysis of linked data once available

Cat. no: PHE 316

Findings from this report:

- The AIHW has established a COVID-19 linked data set
- Over 90% of records supplied for the project were linked
- The project will strengthen evidence-based public health and health system planning for future pandemics
- The linked data can be used to monitor the outcomes and health system use of people who have had COVID-19

Background

This report has been archived. Content previously included in this report can now be found at *COVID-19 register and linked data set* and *COVID-19 linked data set: Linkage results*.

COVID-19 represents an unprecedented health emergency in Australia. The pandemic has led to substantial individual, health system, and broader social and economic effects which will continue to evolve into the future.

Emerging evidence of the medium and long-term effects of COVID-19 presents a need for timely information and monitoring of the health outcomes and health system needs of people who have had a COVID-19 diagnosis.

Collection of data that allow analysis of specific populations, such as those with pre-existing conditions, is vital to informing the community about potential population groups at risk of poorer health outcomes after COVID-19 diagnosis and enabling health systems to identify and manage these cases.

In April 2022, the AIHW was funded $3 million by the Medical Research Future Fund to establish a national linked data platform using existing health data sets to strengthen evidence-based public health and health system planning and management for current and future pandemics.

What is data linkage?

Data linkage is the process of identifying, matching, and merging records that correspond to the same person or entity from several datasets or even within one dataset.

This improves data completeness and provides a rich person-level source of information beyond that available through routine disease surveillance and single data sources. The AIHW uses robust methods to carry out data linkage, and more information can be found at AIHW's data linkage services.

Benefits of linking COVID-19 data

This report has been archived. Content previously included in this report can now be found at *COVID-19 register and linked data set* and *COVID-19 linked data set: Linkage results*.

How has linked data been used during the pandemic?

Linked data is being used internationally to explore a range of important research questions on COVID-19, including:

- characteristics and impact of long COVID (Sivan et al. 2022; Kikkenborg Berg et al. 2022; Murch et al. 2022)
- vaccine uptake and effectiveness (Vasileiou et al. 2022; Perry et al. 2022; Nunes et al. 2021; Mirahmadizadeh et al. 2022)
- health care usage (Lambourg et al. 2022; Lai et al. 2022; Kennedy et al. 2022; Murch et al. 2022; Mirahmadizadeh et al. 2022; Krutikov et al. 2022; Davies et al. 2021)
- risk factors for severe disease (Drefahl et al. 2020; Gao et al. 2021; Liu et al. 2021)
- reinfection rates (Cavanaugh et al. 2021; Mensah et al. 2022).

In Australia, there have been some state-based linkage projects that have been extremely insightful, but a national perspective has been missing from the picture (Rowe et al. 2022; Henry et al. 2021). An additional barrier is that linkage projects can be lengthy and costly for researchers. The Australian COVID-19 linked data set project aims to bridge these gaps.

Why is the project important?

The findings of the project will provide many benefits, including:

- The public will benefit as the data will be used to identify risk factors for severity, long term effects and re-infection. This will then inform service planning and guidelines for the treatment and management of COVID-19 and improve immediate, medium, and longer-term health outcomes for people who have had a COVID-19 diagnosis.
- The public will also benefit from the publication of research findings through fact sheets and reports that will make robust, accessible information available in the public domain.
- Health service providers will be able to gain a better understanding of the service and treatment needs of people who have been diagnosed with COVID-19, and particularly those who develop long term effects of COVID-19 or have had a COVID-19 vaccine, as well as interactions with pre-existing comorbidities.
- Health services providers will understand patterns of service use and medication dispensing by people who have tested positive to COVID-19.
- The research community will benefit from the contribution of knowledge the project will make by filling existing research gaps.
- Existing surveillance systems will be improved through return of linked data at a national level and therefore improve national surveillance reporting.
- The data will enable researchers to monitor and evaluate policies and programs implemented throughout pandemics.
- The platform will provide a foundation for a wide range of future research purposes without these parties having to start from scratch.
- The data will help Commonwealth, state and territory governments plan and manage health resources, for example in aged care facilities.

Aims of the project

This report has been archived. Content previously included in this report can now be found at *COVID-19 register and linked data set* and *COVID-19 linked data set: Linkage results*.

The aims of the project include:

- To establish a national linked data platform that integrates relevant existing health data sets for the purposes of strengthening evidence-based public health and health system planning and management for current and future pandemics.
- To provide an evidence base for epidemiological and statistical research into medium- and long-term effects of a COVID-19 diagnosis.
- To inform health service planning, evaluation, and policy development.
- To inform research projects investigating important health issues such as the impact of vaccination against COVID-19, burden of disease estimates, severity of outcomes for COVID-19 cases, etc.
- To improve the quality of national notifiable disease data.
- To support the states and territories to improve the quality of COVID-19 case data, by returning linked data to custodians.

Data and methods

This report has been archived. Content previously included in this report can now be found at *COVID-19 register and linked data set* and *COVID-19 linked data set: Linkage results*.

Ethics approvals

Before any data could be linked, the project needed to receive ethics approval and funding. The project has 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 NSW 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.

In addition to the ethics approvals outlined above, the data custodian of each state/territory or national dataset also had to approve data usage in line with any jurisdictional requirements.

How was the data linked?

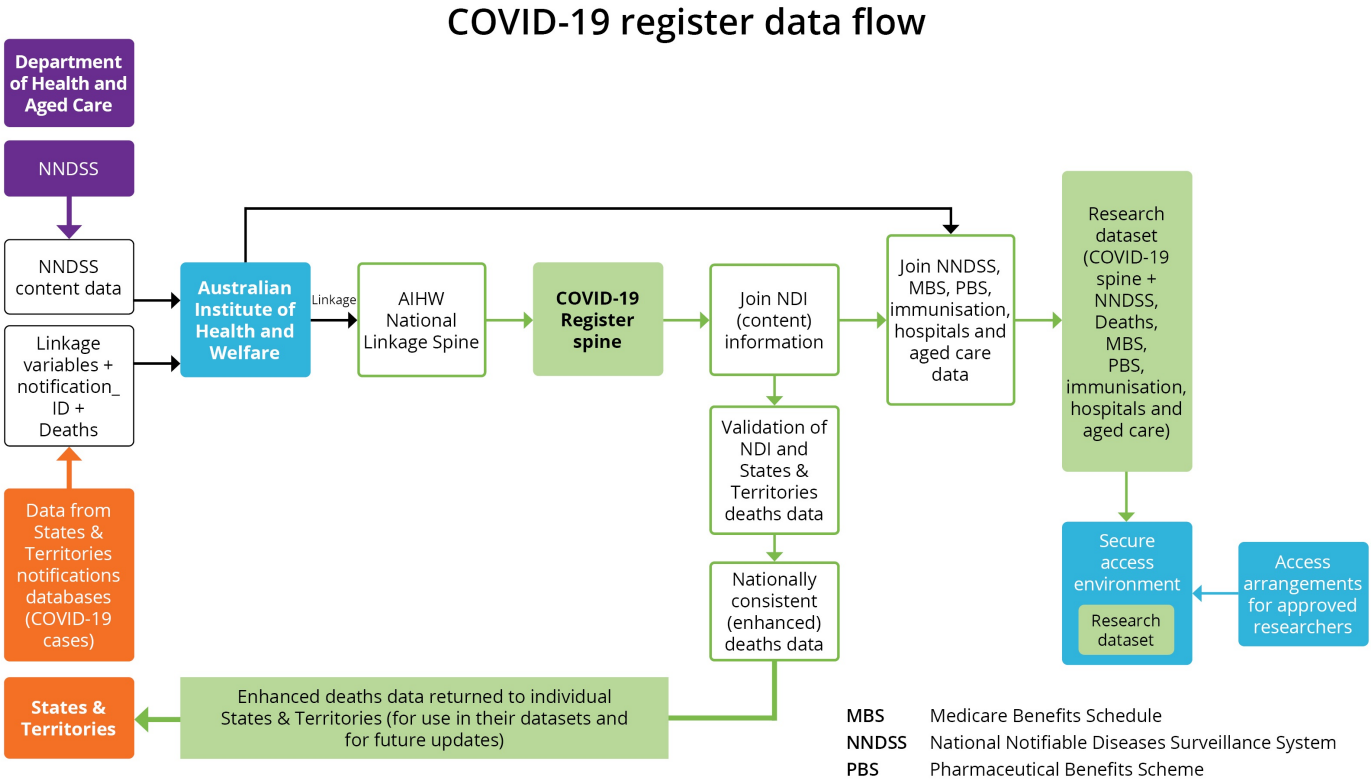
As a Commonwealth Accredited Data Service Provider, the AIHW has the expertise and infrastructure to undertake complex national data linkage.

Linkage variables of the COVID-19 cases were sourced from participating states and territories for linkage purposes. This was then linked with information on AIHW's linkage spine (Medicare Consumer Directory (MCD), National Death Index (NDI) and Australian Immunisation Register (AIR)), using probabilistic record linkage. 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.

Analytical information on COVID-19 cases from states and territories and the Commonwealth Department of Health National Notifiable Disease Surveillance System (NNDSS) has been combined with information from the NDI, Medicare Benefits Schedule (MBS) and Pharmaceutical Benefits Scheme (PBS, including Repatriation Schedule of Pharmaceutical Benefits (RPBS) information), the National Hospitals Morbidity Database (NHMD), the National Non-Admitted Patient Emergency Department Care Database (NNAPEDCD), the National Aged Care Data Clearinghouse (NACDC) and the AIR to create a de-identified linked research data set. Figure 1 outlines the linkage processes for the current project.

The AIHW data linkage protocols prescribe strict separation of identifiers and analytical data within the AIHW linkage team, so that where staff have access to personal identifiers and analytical data for study participants, they will not have access to the identifiers and analytical data at the same time for the duration of the project.

Figure 1. COVID-19 linked data flow



How often will the data be updated?

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

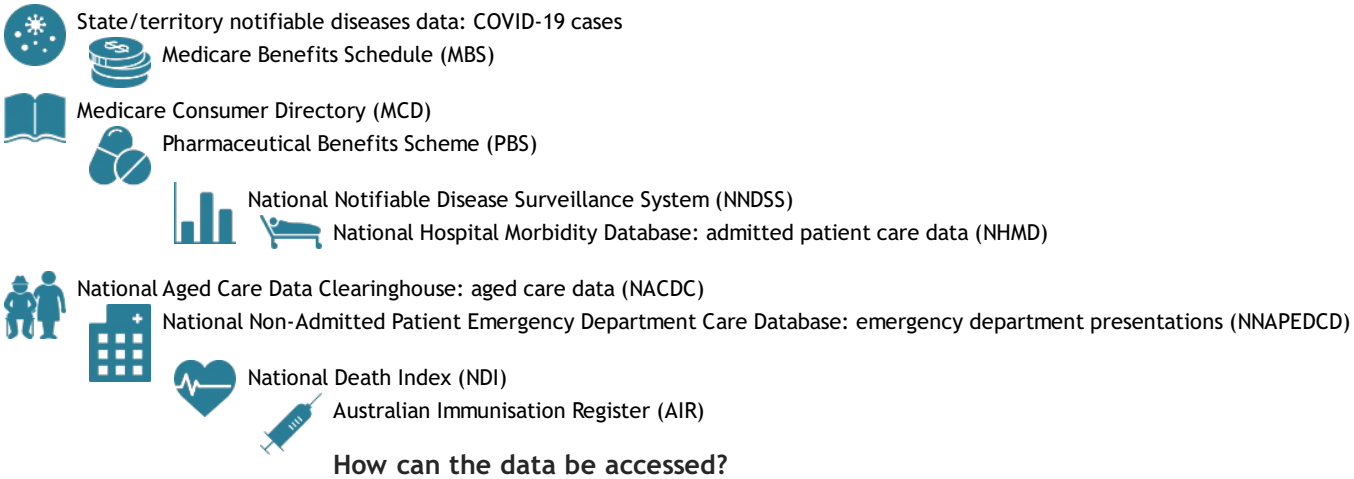
How are linked data being returned to states and territories?

After both the initial and re-linkages, date of death and cause of death information from the NDI will be released to the states and territories that provided 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.

What data sets are included?

Data sets available as of December 2022 are listed in Figure 2 below. Future iterations of the project will look to add additional sources of information.

Figure 2: Figure 2: COVID-19 linked datasets available as of March 2023



How can the data be accessed?

All users who want to access the de-identified research data will be required to submit to AIHW a project proposal including a data analysis plan and a signed Australian Institute of Health and Welfare Act 1987 s29 Undertaking of Confidentiality form. This form protects the privacy of individuals by making it a criminal offence to disclose information about the participants of a study, punishable by fines and/or imprisonment. Data will not be provided to, accessed, or used by another, unauthorised party. Access is strictly controlled within a secure remote access environment, with no access allowed to other project workspaces.

Due to the detailed and sensitive nature of the data, access will only be provided via secure research environments where AIHW can apply appropriate vetting and management processes in line with AIHW's Five Safes Framework.

In this first stage of the project, only government researchers or those funded by government 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 and a suitable secure environment is available.

Initial linkage findings

This report has been archived. Content previously included in this report can now be found at *COVID-19 register and linked data set* and *COVID-19 linked data set: Linkage results*.

Scope of the data

The linked data set currently includes data provided to the AIHW by the following jurisdictions:

- Australian Capital Territory
- New South Wales
- Northern Territory
- South Australia
- Tasmania.

This is the first iteration of the data, which will be regularly updated, with the aim to include all jurisdictions and more data sets in future iterations. Data have been received from Queensland and Victoria and will be available in the next iteration.

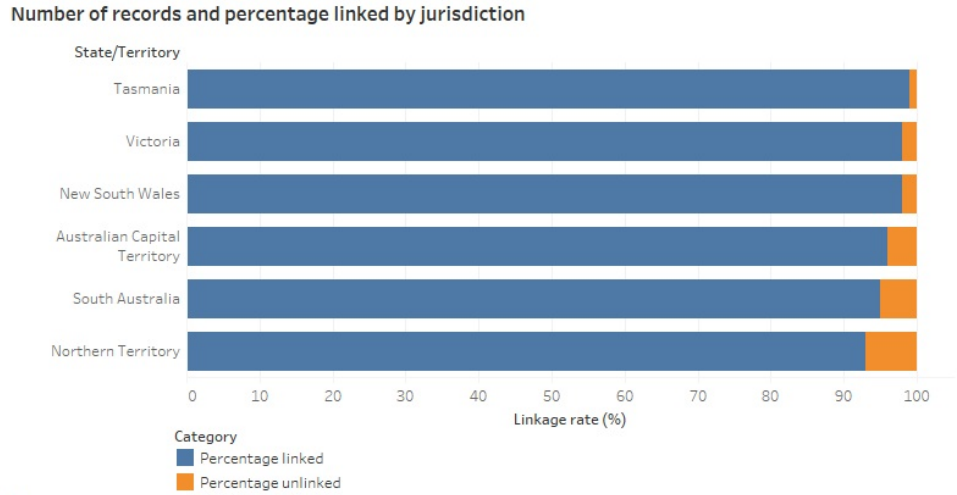
Linkage rates by jurisdiction

COVID-19 case linkage variables (names, addresses, dates of birth and sex) provided by jurisdictions to AIHW were probabilistically linked to AIHW National Linkage Spine (NLS). AIHW NLS combines linkage variables from MCD, AIR and NDI and covers almost all of the population of Australia. The linkage results depend on the accuracy and completeness of the linkage variables provided to AIHW: more accurate and complete data result in better linkage rates.

Figure 3 shows the number of records that were linked and those that were unable to be linked by state and territory. For all jurisdictions, over 90% of records supplied for the project were linked. The lower linkage rate in the Northern Territory may be due to limited address information provided with the case data. AIHW is working with the Northern Territory to improve this rate.

Figure 3: Number of records and percentage linked by jurisdiction

The segmented horizontal bar chart shows Tasmania has the highest percentage of linked records (99%) and Victoria has the highest number of individuals linked (2,536,790 people) in the first iteration of the COVID-19 linked data set. All jurisdictions have over 90% of records linked.



Note:
Initial linkage for Victoria is complete and will be available in the next release. All other jurisdictions available for analysis from December 2022.
Source:
http://www.aihw.gov.au

Linkage rates by population groups

Table 1 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 doing analysis on linked data. For example, individuals who change addresses whilst renting may also be underrepresented in linkage studies. Table 1 shows that for all groups except the 'Other' sex/gender category, over 90% records were linked. Sex is one of the key variables used to link records, therefore, where sex is not reported consistently, or as male or female ('Other' in table below) linkage rates are lower. Individuals aged 70+ had the highest percentage of records unlinked (9.9%), however there were not large differences in linkage rates across the age groups.

Table 1. Linkage rates by population groups

	No. of records linked (%)	No. of records not linked (%)
Sex/gender[1]		
Male	1,305,834 (97.7%)	30,838 (2.3%)
Female	1,472,188 (97.9%)	32,111 (2.1%)
Other[2]	9,583 (75.6%)	3,097 (24.4%)
Age group		
0-15	505,626 (97.4%)	13,421 (2.6%)
16-29	655,905 (97.7%)	15,513 (2.3%)
30-49	923,626 (98.8%)	11,645 (1.2%)
50-69	517,968 (99.0%)	5,152 (1.0%)
70+	184,480 (90.1%)	20,203 (9.9%)

1. As reported by the state and territory.
2. Other includes records where sex or gender is not reported, or sex is reported as neither male nor female.

Future developments

This report has been archived. Content previously included in this report can now be found at *COVID-19 register and linked data set* and *COVID-19 linked data set: Linkage results*.

The next steps of this project are to improve the coverage of COVID-19 cases, which includes linking more recent case data and linking the remaining jurisdictions' case information.

As more recent information from the other data sources becomes available e.g., hospitals data, this will be reflected in the linked data as well. Updating the content data regularly will provide a growing longitudinal resource for the original cases, to follow their health journey over time.

Future stages of this project will look at linking other sources of information, such as the National Disability Insurance Scheme. This will ensure the information can look at the impacts of COVID-19 in other sectors outside the health system.

To ensure value for money and improve efficiencies, the AIHW research team are keeping abreast of progress in the data linkage space and seek to align with, and make use of, cloud-based national platforms for data sharing, linking and access as they develop. The linked data platform has been built to ensure interoperability with existing national data sets that have been used during the pandemic, such as the Census.

References

Cavanaugh A. M., Spicer K. B., Thoroughman D, Glick C, Winter K (2021) 'Reduced Risk of Reinfection with SARS-CoV-2 After COVID-19 Vaccination — Kentucky, May–June 2021'. *MMWR Morb Mortal Wkly Rep*, 70:1081-1083,doi: doi: 10.15585/mmwr.mm7032e1.

Davies A, Song J, Akbari A, Bentley L, Carter B, Cross L, Dundon J, Florentin D, Newman C, Smith T, Trigg L, John G (2021) 'Impact of the COVID-19 pandemic on health-care usage and mental health of clinically extremely vulnerable individuals in Wales: a population-scale data linkage study.' *The Lancet* 398(S39), doi: 10.1016/S0140-6736(21)02582-4.

Drefahl S, Wallace M, Mussino E, Aradhya S, Kolk M, Branden, Malmberg B, Andersson G (2020) 'A population-based cohort study of socio-demographic risk factors for COVID-19 deaths in Sweden'. *Nat Commun,* 11, 5097. doi: 10.1038/s41467-020-18926-3.

Gao M, Piernas C, Astbury N. M., Hippisley-Cox J, O'Rahilly S, Aveyard P, Jebb S. A. (2021) 'Associations between body-mass index and COVID-19 severity in 6.9 million people in England: a prospective, community-based, cohort study'. *The Lancet Diabetes & Endocrinology* 9:350-9. doi: 10.1016/S2213-8587(21)00089-9.

Kikkenborg Berg S, Palm P, Nygaard U, Bundgaard H, Petersen M, Rosenkilde S, Thorsted A. B., Ersbøll A. K., Thygesen L. C., Nielsen S. D., & Vinggaard Christensen, A (2022) 'Long COVID symptoms in SARS-CoV-2-positive children aged 0–14 years and matched controls in Denmark (LongCOVIDKidsDK): a national, cross-sectional study'. *The Lancet Child & Adolescent Health*, 6(9):614–623. doi: 10.1016/S2352-4642(22)00154-7.

Liu B, Spokes P, He W, Kaldor J (2021) 'High risk groups for severe COVID-19 in a whole of population cohort in Australia'. *BMC Infect Dis* 21, 685. doi: 10.1186/s12879-021-06378-z.

Kennedy J, Parker M, Seaborne M, Mhereeg M, Walker A, Walker V, Denaxas S, Kennedy N, Katikireddi V.S, Brophy S (2022) 'Health care use attributable to COVID-19: A propensity matched national electronic health records cohort study of 249,390 people in Wales, UK' *medRxiv* doi: 10.1101/2022.04.21.22274152v1.

Krutikov M, Stirrup O, Nacer-Laidi H, Azmi B, Fuller C, Tut G, Palmer T, Shrotri M, Irwin-Singer A, Baynton V, Hayward A, Moss P, Copas A, Shallcross L, The COVID-19 Genomics UK consortium (2022) 'Outcomes of SARS-CoV-2 omicron infection in residents of long-term care facilities in England (VIVALDI): a prospective, cohort study' *The Lancet Healthy Longevity*, 3(5): E347-E355, doi: 10.1016/S2666-7568(22)00093-9.

LaiF.T.T., Huang L, Peng K, Li X, Chui C.S.L, Wan E.Y.F, Wong C.K.H, Chan E.W.Y, Hung I.F.N, Wong I.C.K (2022)' Post-Covid-19-vaccination adverse events and healthcare utilization among individuals with or without previous SARS-CoV-2 infection' *Journal of Internal Medicine*, 291(6):864-869, doi: 10.1111/joim.13453.

Lambourg E, Gallacher P.J, Hunter R.W, Siddiqui M, Miller-Hodges E, Chalmers J, Pugh D, Dhaun N, Bell S (2022) 'Cardiovascular outcomes in patients with chronic kidney disease and COVID-19: a multi-regional data-linkage study', *European Respiratory Journal*, doi: 10.1183/13993003.03168-2021.

Mensah A. A., Lacy J, Stowe J, Seghezzo G, Sachdeva R, Simmons R, Bukasa A, O'Boyle S, Andrews N, Ramsay M, Campbell H, Brown K (2022) 'Disease severity during SARS-COV-2 reinfection: a nationwide study'. *Journal of Infection* 84:542-50. doi: 10.1016/j.jinf.2022.01.012.

Mirahmadizadeh A, Heiran A, Lankarani K.B, Serati M, Habibi M, Eilami O, Heiran F, Moghadami M (2022) 'Effectiveness of Coronavirus Disease 2019 Vaccines in Preventing Infection, Hospital Admission, and Death: A Historical Cohort Study Using Iranian Registration Data During Vaccination Program'. *Open Forum Infectious Diseases*, 9(6):ofac177, doi: 10.1093/ofid/ofac177.

Murch B, Hollier S.E, Kenward C, Wood R.M (2022) 'Use of linked patient data to assess the effect of Long-COVID on system-wide healthcare utilisation', *Health Information Management Journal*18333583221089915, doi: 10.1177/18333583221089915.

Nunes B, Rodrigues A.P, Kislaya I, Cruz C, Peralta-Santos A, Lima J, Leite P.P, Sequeira D, Dias C.M, Machado A (2021) 'mRNA vaccine effectiveness against COVID-19-related hospitalisations and deaths in older adults: a cohort study based on data linkage of national health registries in Portugal, February to August 2021' *Eurosurveillance*, 26(38): 2100833.

Perry M, Gravenor M.B, Cottrell S, Bedston S, Roberts R, Williams C, Salmon J, Lyons J, Akbari A, Lyons R.A, Torabi F, Griffiths L.J (2022) '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*, 18(1):2031774, doi: 10.1080/21645515.2022.2031774.

Sivan M, Greenhalgh T, Darbyshire J.L, Mir G, O'Connor R.J, Dawes H, Greenwood D, O'Connor D, Horton M, Petrou S, de Lusignan S, Curcin V, Mayer E, Casson A, Milne R, Rayner C, Smith N, Parkin A, Preston N, Delaney B (2022) 'LOng COvid Multidisciplinary consortium Optimising Treatments and servIces acrOss the NHS (LOCOMOTION): protocol for a mixed-methods study in the UK' *BMJ Open*,12(5):e063505, doi: 10.1136/bmjopen-2022-063505.

Vasileiou E, Shi T, Kerr S, Robertson C, Joy M, Tsang R, McGagh D, Williams J, Hobbs R, de Lusignan S, Bradley D, OReilly D, Murphy S, Chuter A, Beggs J, Ford D, Orton C, Akbari A, Bedston S, Davies G, Griffiths L.J, Griffiths R, Lowthian E, Lyons J, Lyons R.A, North L, Perry M, Torabi F, Pickett J, McMenamin J, McCowan C, Agrawal U, Wood R, Stock S.J, Moore E, Henery P, Simpson C.R (2022) Investigating the uptake, effectiveness and safety of COVID-19 vaccines: protocol for an observational study using linked UK national data', *BMJ Open*,12(2):e050062, doi:10.1136/bmjopen-2021-050062.

Related material

Resources