CSIRO PUBLISHING

Australian Health Review
https://doi.org/10.1071/AH18014

Recent trends in the use of linked data in Australia

Angela Young[1] BSc(Hons), PhD, Coordinator Policy and Client Services

Felicity Flack[1,2] BSc(Hons), PhD, Manager Policy and Client Services

[1]Research Infrastructure Centres, The University of Western Australia. M320, 35 Stirling Highway Crawley, WA 6009, Australia. Email: angela.young@uwa.edu.au

[2]Corresponding author. Email: felicity.flack@uwa.edu.au

Abstract

Objective. The aim of this study was to quantify the use of linked data for health and human services research in Australia since the establishment of the Population Health Research Network (PHRN) in 2009.

Methods. A systematic literature search was performed using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2009 checklist to search for all publications involving the use of Australian linked data between 2009–10 and 2016–17. Publications were categorised by subject, data linked and data linkage unit involved.

Results. In all, 7153 articles were identified from the initial search, and 1208 were included in the final analysis. An increase in the number of publications involving linked data was observed from 2009–10 through to 2015–16. Most articles (82%) featured data linked by at least one PHRN-funded data linkage unit. The research areas of 86% of publications were able to be classified according to the International Statistical Classification of Diseases and Related Health Problems 10th Revision Australian Modification (ICD-10-AM). The number of publications involving cross-sectoral linked data also increased.

Conclusions. Investment in Australian data linkage infrastructure has seen an increase in the number of research publications involving the use of linked health and human services data. This study identified areas where linked data is commonly used and those where use could be improved.

What is known about the topic? Data linkage is a method of bringing together information about individual people, places and events from different sources in a way that protects individual privacy. Individual jurisdictions have reported benefits from research conducted using linked data, including the generation of new knowledge and supporting improvements in the delivery of a wide range of health and human services. There has been significant investment in national data linkage infrastructure in Australia over the past 8 years. To date, there has been no systematic investigation of the effect of this investment on the use of linked population data by the research community.

What does this paper add? This paper provides evidence of the increased use of high-quality population-based linked data in research over the 8-year period studied. It demonstrates the application of data linkage across a wide range of health areas and highlights the small but growing number of studies using cross-sectoral data to investigate complex conditions.

What are the implications for practitioners? It is important to demonstrate to funders, policy makers, data custodians and researchers the value of robust data linkage capacity as an important national resource. Its use by researchers can bring enormous social and economic benefits by providing a more complete picture of the health and well-being of the community. The range of data collections routinely linked is increasing, as is the pool of researchers experienced in handling and analysing the data. Continued investment in Australia's data linkage infrastructure and the inclusion of other collections including general practice data will augment the use of this infrastructure in expanding the evidence base for policy makers and practitioners.

Received 16 January 2018, accepted 22 May 2018, published online 27 August 2018

Introduction

Data linkage is a technique for creating links between data from different sources that relate to the same person, family, place or event. This technique enables researchers to conduct studies on populations using linked data from a range of different data sources.[1] Australia has been at the forefront of the development of data linkage methods to provide researchers access to linked data while preserving privacy. The Western Australian Data Linkage System (WADLS) was established in 1995 and the Centre for Health Records Linkage (CHeReL) was established in New South Wales (NSW) and the Australian Capital Territory (ACT) in 2006.[2,3] The Australian Institute of Health and Welfare (AIHW), a Commonwealth statutory authority that reports on

Australia's health and welfare, has linked data from a range of administrative data collections, including the National Death Index and the Australian Cancer Database, since the mid-1990s for its own work and on behalf of clients.[4]

Since 2009, the Australian Government has made a substantial investment in the development and operation of national data linkage infrastructure through the Population Health Research Network (PHRN).[5] A total of A$55 million has been provided to the PHRN by the Australian Government through the National Collaborative Research Infrastructure Strategy and related programs to mid-2019. State and territory governments and academic partners have also made considerable investments. This infrastructure has the capacity to support research of national and international importance.

There is limited published research that has documented the extent and effect of data linkage infrastructure in Australia. The achievements of the WADLS and CHeReL have been detailed and output measures, such as the number of approved projects that have received linked data, the number of publications that have resulted from the use of the infrastructure, the number of research degrees completed and the number of records included in the Master Linkage Key, have been used to describe these achievements.[2,3,6,7] A recent systematic review of linked hospital data showed that the number of these publications has been increasing since the late 1990s.[8]

Linkage of the Commonwealth's Pharmaceutical Benefits Scheme (PBS) data to other administrative or cohort data collections provides an opportunity to conduct post-market drug surveillance or the evaluation of targeted drug interventions. Historically, access to PBS data for linkage studies has been restricted to those involving informed consent, with the exception of a small number of studies undertaken through the WADLS where a time-limited memorandum of understanding for use of PBS data with a waiver of consent was in place.[2]

The objective of the present study was to quantify the use of linked data for health and human services research in Australia since the establishment of the PHRN in 2009. The study also evaluated whether researchers' access to PBS data had improved in this time frame. The analysis focuses on the number of peer-reviewed publications and the research areas studied.

Methods

Information sources and search strategy

The literature search strategy was based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2009 guidelines.[9] The electronic databases Ovid MEDLINE, PsycINFO, Embase, EconLit and Scopus were systematically searched between March and July 2017 to identify all the relevant articles published between 1 July 2009 and 30 June 2017. The search terms used were data*, record*, link* and Australia. The search was limited to English-language, peer-reviewed journal articles and human studies.

The PHRN,[10] CHeReL[11] and BioGrid[12] websites all provide lists of publications arising from the use of their data linkage infrastructure. All these sources were also searched for articles that met the search criteria of the present study.

Study selection and eligibility criteria

Results were compiled using reference manager software (Endnote X8, Clarivate Analytics, Philadelphia, PA, USA) and duplicates were removed. The authors independently screened titles and abstracts against the inclusion and exclusion criteria. To be eligible for inclusion, the studies had to be published in the English language, be human studies, peer-reviewed articles and to involve the use of linked Australian data, defined as linkage of two or more data collections at the unit record level. Methods of linkage included, but were not limited to, manual and the use of a specialist data linkage unit. Registries were counted as a single data collection. Articles that did not involve the use of Australian linked data, covered data linkage methods and linkage software evaluation and were conference abstracts, reviews, comments and letters were excluded.

Full-text screening was performed where it was not possible to determine from the title or abstract whether linked data were used. Discrepancies were resolved by agreement between the two reviewers.

Classification

The authors screened the full text of all the eligible articles and classified them according to the following criteria:

- financial year (the date of publication was defined as the first time the paper was made publicly available, whether online or in print)
- data linkage unit (PHRN unit or other, which includes articles where the unit could not be determined)
- research area according to the International Statistical Classification of Diseases and Related Health Problems 10th Revision Australian Modification (ICD-10-AM)[13]
- research area according to national health priority area (NHPA)[14]
- study used linked PBS data (yes/no)
- study involving cross-sectoral linked data.

The identity of the unit that conducted the data linkage was not reported in some publications. In these cases the authors consulted with the PHRN data linkage units located in the jurisdiction(s) of the data sources to attribute the publication's linkage to the correct unit.

The data were then extracted into a Microsoft (Bellevue, WA, USA) Excel spreadsheet for analysis.

Results

From the systematic literature search, 7153 articles were identified from the initial search, 2773 articles were duplicates identified from the different databases and 819 were not peer-reviewed journal articles. The remaining 3561 articles were screened against the inclusion and exclusion criteria, and 2353 articles were excluded, leaving 1208 articles for inclusion in the final analysis.

The number of publications using linked data in Australia has been increasing since 2009–10 (Fig. 1), with a peak in 2015–16 followed by a slight decline in 2016–17. The linkage for the majority of articles ($n = 987$; 82%) was performed by one or more PHRN data linkage units (Fig. 1). PHRN data linkage units are

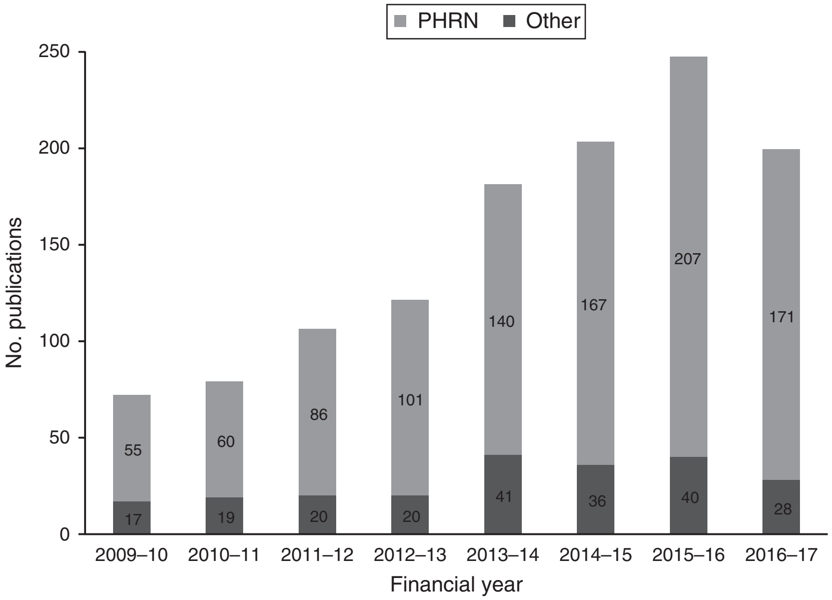

Fig. 1. Total number of peer-reviewed publications involving linked data in Australia classified by the use of data linked by a Population Health Research Network (PHRN)-funded data linkage unit versus other linkage unit, from 2009–10 to 2016–17.

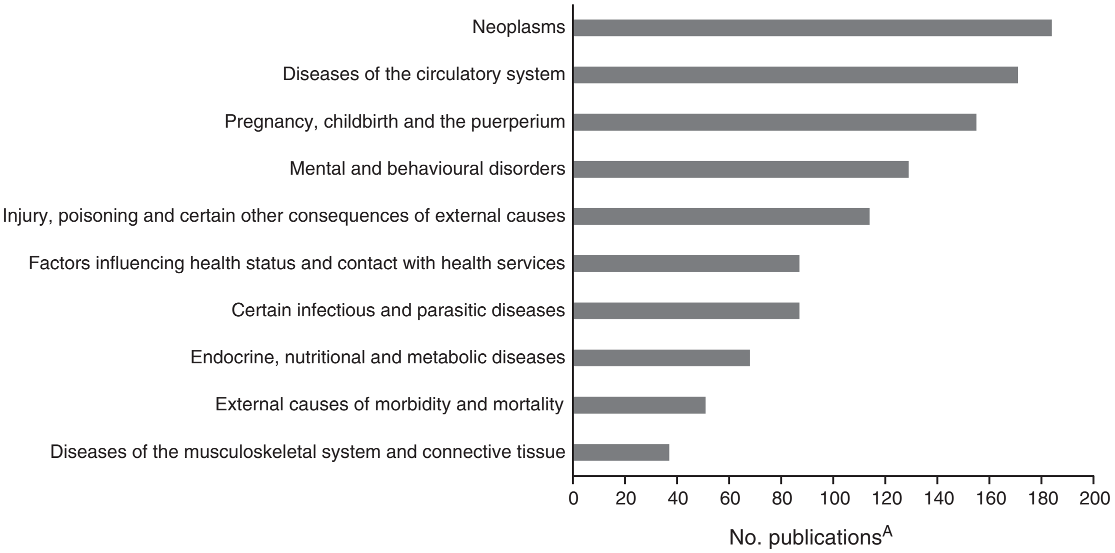

Fig. 2. The top 10 International Statistical Classification of Diseases and Related Health Problems 10th Revision Australian Modification (ICD-10-AM) chapters addressed in peer-reviewed publications involving linked data in Australia, from 2009–10 to 2016–17. ^{A}Note, publications may be counted more than once if they span more than one ICD-10-AM chapter.

defined as specialist data linkage units that have ever received funding from the PHRN. Non-PHRN data linkage is defined as linkage conducted by the researcher, a data custodian or a specialist data linkage service not funded by the PHRN. The proportion of publications where the linkage was conducted by PHRN data linkage units has also increased from 79% in 2009–10 to 86% in 2016–17.

The research areas covered in most publications were related to health. The publications were classified according to the ICD-10-AM, and 88% or 1058 of those identified mapped to at least one ICD-10-AM chapter. The 10 most common research areas (in descending order) were determined as follows (Fig. 2): neoplasms; diseases of the circulatory system; pregnancy, childbirth and the puerperium; mental health and behavioural disorders; injury, poisoning and certain other consequences of external causes; factors influencing health status and contact with health services; certain infectious and parasitic diseases; endocrine, nutritional and metabolic diseases; external causes of morbidity and mortality; and diseases of the musculoskeletal system and connective tissue.

Publications were also categorised on the basis of their research focus with regard to the nine identified NHPAs. Of the 1208 publications in total, 681 (56.3%) addressed at least one of the NHPAs, with 50 of these (7.3%) covering two or more. The most represented of the NHPAs were cancer control, cardiovascular health, injury prevention and control and mental health (Fig. 3).

The use of linked data for research involving an analysis of prescribed pharmaceuticals was assessed by counting the number of publications reporting the use of linked PBS data. PBS data contain information on all dispensing events of medicines available to patients at a government-subsidised price and is managed by the Department of Human Services. In 2009–10 there was only one study captured in our analysis that involved the linkage of PBS data to other collections, representing 1% of all studies using linked data in this period. The following financial year saw an increase to seven publications, or 9% of linked data studies, involving PBS data. This level has been approximately maintained through to 2016–17, when there were 17 studies (8%) using linked PBS data (data not shown).

Although the majority of publications focused specifically on health, some studies included analysis of data from other sectors, such as education, justice and child protection. The number of publications involving data from multiple sectors was counted. In 2009–10, 11 publications (15.3%) involved cross-sectoral linkage. This proportion remained relatively consistent in 2016–17 (14.3%), but the total number of cross-sectoral publications for this period had risen to 28 (Fig. 4).

Discussion

In the past decade there has been increasing interest in the use of linked administrative data for research in Australia. Establishment of the PHRN with support from the Australian Federal, state and territory governments and academic institutions has led to a significant expansion of the nation's data linkage infrastructure. This has enabled Australian researchers to undertake innovative, world-class research previously not possible. It is important to evaluate how this publicly funded infrastructure has been used and to demonstrate the benefits to the community.

A previous report on the growth of linked hospital data use in Australia indicated that there have been improvements in access to linked data since 1995.[8] The present study is the first to demonstrate the growth of the use of linked data in health and human service research outputs in Australia from 2009–10 to 2016–17.

Over the study period (2009–17), a nearly 3.5-fold increase in the number of publications using linked data was observed. This increase coincides with the development of the PHRN. At the time of its establishment in 2009, only two states and one territory had access to a data linkage unit. In 2017, every state and territory had access to a data linkage unit and there was a significant expansion in national data linkage capability through the AIHW.

The slight decline in publication numbers observed in 2016–17 from the previous reporting year may be due to a limitation in methodology. The literature search was conducted at a single point in time and the date limits used in the database search identified articles by their journal publication date, rather than the date they first appeared online (which, for many journals, often predates the former). As a result, it is predicted that several articles that were available in the public domain at the time of the search but had not been formally published in a journal, and were therefore not yet captured in the online databases searched, would have been missed, particularly for this most recent period.

The upward trajectory in the total number of publications was paralleled by an increase in the proportion of publications using data linked by PHRN data linkage units. The robust data governance processes used by these data linkage units are likely to

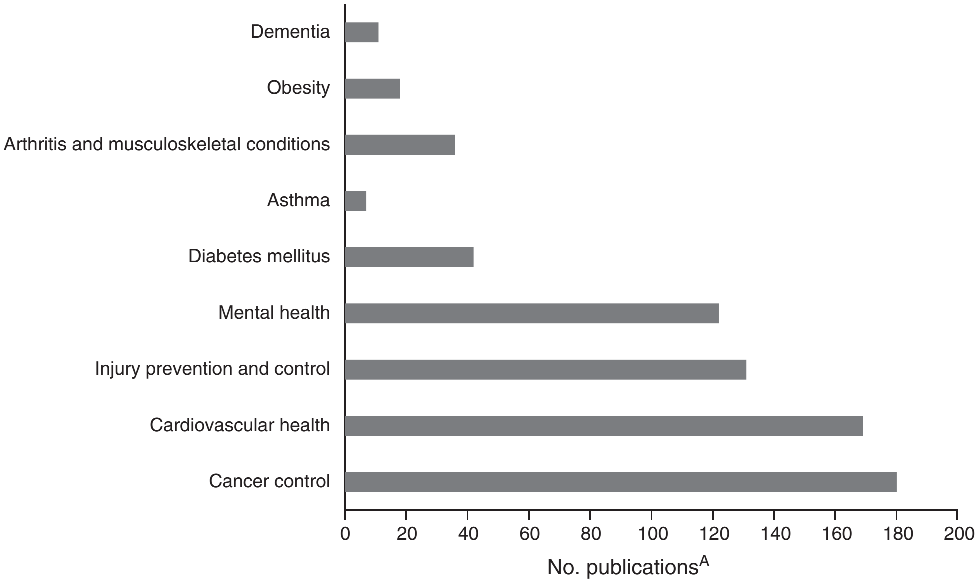

Fig. 3. Peer-reviewed publications addressing at least one national health priority area (NHPA), stratified by NHPA category, from 2009–10 to 2016–17. ^ANote, publications may be counted more than once if they span more than one NHPA category.

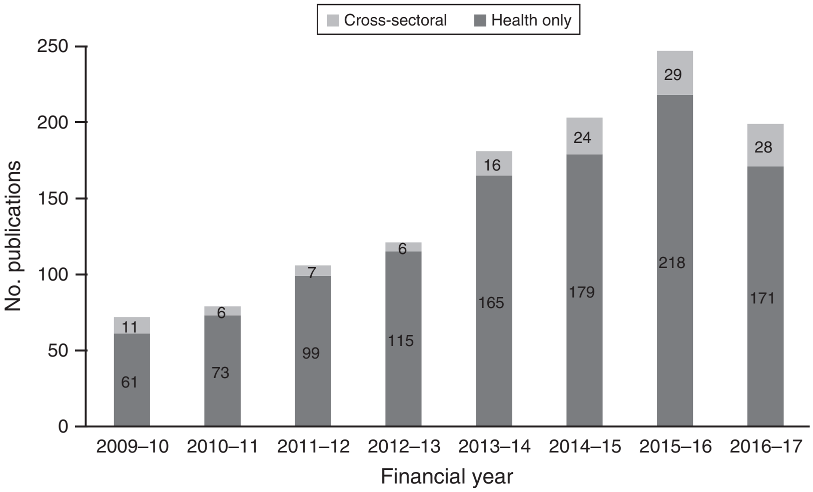

Fig. 4. Total number of peer-reviewed publications classified by those involving cross-sectoral linked data versus health-only data in Australia, from 2009–10 to 2016–17.

have made it more attractive for data custodians to make their data available for linkage rather than directly to researchers on an *ad hoc* basis. All states and territories now have their core health data collections linked, including at least 10 years of hospital data that are linked on an enduring basis.[15]

The nine NHPAs represented those that contribute most to the burden of illness in the Australian population, particularly if the burden could be significantly reduced.[16] Therefore they are a target for research and associated funding. Accordingly, the National Health and Medical Research Council annually directs the bulk of its research funding towards these areas.[17] It was therefore not surprising that the NHPAs featured prominently in data linkage publications published through 2009–17, with 57% of publications in 2016–17 addressing at least one of these focus areas.

Cancer control, cardiovascular disease, injury prevention and control, mental health and diabetes were represented in the majority of data linkage studies involving at least one NHPA. Many of these studies relied on the use of hospital admissions, emergency department and mortality information that, in general, have comprehensive administrative databases available and that are, in most cases, linked on a routine basis by state or territory data linkage units. Several disease-specific data sources, including state or territory cancer registries, mental health collections, trauma registries, ambulance data and the National Diabetes Service Scheme collection, also formed a critical part of the analysis in many of these studies.

Areas of national health priority that have not been the subject of widespread research involving linked data include asthma and other respiratory diseases, obesity, arthritis and musculoskeletal conditions and dementia. One reason for this could be the lack of any jurisdictional primary care data collections or a national minimum primary care data collection available to be linked, because many of these conditions would be managed at the primary care level.[18]

In association with the NHPAs, data linkage publications identified in the present study were also stratified based on ICD-10-AM coding where possible. In general, the most common ICD-10-AM chapters aligned to the most commonly observed NHPAs. There was also a parallel in our observations with those reported through a study of research outputs by WADLS published in 2007.[6] In addition, two ICD-10-AM chapters that featured prominently in the analysis but were not included in the NHPAs were ‘Certain infectious and parasitic diseases’ (A00–B99) and ‘Pregnancy, childbirth and the puerperium’ (O00–O99). Several jurisdictional data linkage units include routinely linked data in their master linkage files that are particularly useful for conducting investigations in these areas, including notifiable disease databases, pathology data and immunisation data, as well as perinatal data.

The analysis showed an initial jump in publications including linked PBS data from 1% of publications in 2009–10 to 9% in 2010–11. This is due to publications using linked data through the WADLS–Commonwealth agreement. The percentage of publications using linked PBS data has remained relatively consistent since this time. Recent changes in data governance and integration arrangements within the Commonwealth,[19] as well as PHRN initiatives to improve the linkage of Commonwealth to state and territory data, are expected to increase the availability of PBS data for access by the linked data research community, leading to higher numbers of research outputs using these data.

Analysis of publications resulting from the use of linked data in Australia has provided evidence that it is being used more widely by the research community for health research. An increasing number of publications was identified involving the cross-sectoral linkage of health data with collections such as education, justice, police, child protection and environmental data. Several PHRN-funded data linkage units already have a range of non-health data collections in their master linkage files,[20–22] and it is expected that this number will grow in the effort to address many of the multidimensional ‘wicked problems’ that exist in Australia.[23]

Despite the strong growth in publications from research using linked data over the past 8 years, several impediments remain to timely access to this data.[24,25] The PHRN is actively working with its funded linkage units and other network participants, data custodians and the research community to streamline access and reduce the time and costs of data linkage and supply. A recent PHRN survey of Australian linked data researchers has identified collections that are considered of high value, and efforts will be directed at making these data more readily available for inclusion in data linkage systems wherever possible.[26] It is also expected that the implementation of reforms recommended in the recent Australian Government's Productivity Commission Inquiry Report into Data Availability and Use[24] will be an important driver in harnessing the power of linked data for vital health and related research.

Limitations

Given that our search terms required the keyword 'link*' to include articles in the output, those papers that utiLised data linkage and/or the services of a linkage unit but did not articulate this clearly would not have been identified, leading to a presumed level of underreporting. A recent international collaboration has seen the development of reporting guidelines that aim, among other things, to increase the discovery of publications involving the use of routinely collected data including in the linkage setting.[27] The promotion and adoption of these guidelines by the research community would assist in overcoming this underreporting issue.

Conclusions

Investment in Australian data linkage infrastructure has resulted in a steady increase in the number of research publications involving the use of linked health and human services data. The present study has highlighted areas where linked data is commonly used and areas where its use could be improved, for example in the inclusion of general practice data, to facilitate world-class research into asthma, obesity, arthritis and dementia. There has been an increase in publications based on research using cross-sectoral linked data, and this will help inform evidence-based approaches to some of the complex problems facing Australian society. The PHRN is focused on expanding the number of data collections included in the national data linkage system and providing the necessary support services and infrastructure to enable timely and efficient access to linked data from a wider range of areas and enabling its full potential to be realised.[28]

Further analysis of the types of data collections and research questions may assist in informing the future development of Australia's data linkage infrastructure.

Competing interests

None declared.

Acknowledgements

This study was supported by the Population Health Research Network (PHRN), which is an initiative of the Australian Government being conducted as part of the National Collaborative Research Infrastructure Strategy. The authors thank the PHRN's regional and national data linkage units for their assistance with providing details of completed linked data projects and relevant outputs. The authors also acknowledge Merran Smith for assisting with manuscript review.

References

1 Hobbs MS, McCall MG. Health statistics and record linkage in Australia. *J Chronic Dis* 1970; 23: 375–81. doi:10.1016/0021-9681(70)90020-2

2 Holman CD, Bass AJ, Rosman DL, Smith MB, Semmens JB, Glasson EJ, Brook EL, Trutwein B, Rouse IL, Watson CR, de Klerk NH, Stanley FJ. A decade of data linkage in Western Australia: strategic design, applications and benefits of the WA data linkage system. *Aust Health Rev* 2008; 32: 766–77. doi:10.1071/AH080766

3 Irvine KA, Moore EA. Linkage of routinely collected data in practice: the Centre for Health Record Linkage. *Public Health Res Pract* 2015; 25: e2541548. doi:10.17061/phrp2541548

4 Australian Institute of Health and Welfare. Data linkage. Last updated 8 February 2018. Available at: https://www.aihw.gov.au/our-services/data-linkage [verified 16 May 2018].

5 Population Health Research Network. Population Health Research Network Overview. 2018. Available at: http://www.phrn.org.au/about-us/overview/ [verified 16 May 2018].

6 Brook EL, Rosman DL, Holman CD. Public good through data linkage: measuring research outputs from the Western Australian Data Linkage System. *Aust N Z J Public Health* 2008; 32: 19–23. doi:10.1111/j.1753-6405.2008.00160.x

7 Irvine KA, Taylor LK. The Centre for Health Record Linkage: fostering population health research in NSW. *N S W Public Health Bull* 2011; 22: 17–18. doi:10.1071/NB10061

8 Tew M, Dalziel KM, Petrie DJ, Clarke PM. Growth of linked hospital data use in Australia: a systematic review. *Aust Health Rev* 2017; 41: 394–400. doi:10.1071/AH16034

9 Moher D, Liberati A, Tetzlaff J, Altman DG. The PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ* 2009; 339: b2535. doi:10.1136/bmj.b2535

10 Population Health Research Network. Research publications. 2016. Available at: http://www.phrn.org.au/publications/research-publications/ [verified 16 May 2018].

11 Centre for Health Record Linkage. Publications. 2018. Available at: http://www.cherel.org.au/publications [verified 16 May 2018].

12 Biogrid Australia. Biogrid associated journal publications by year. 2018. Available at: https://www.biogrid.org.au/page/19/journal-publications [verified 16 May 2018].

13 Australian Consortium for Classification Development. The international statistical classification of diseases and related health problems tenth revision Australian modification (ICD-10-AM/ACHI/ACS). Darlinghurst: Independent Hospital Pricing Authority; 2017.

14 Australian Institute of Health and Welfare. First report on the national health priority areas, full report. 1997. Available at: https://www.aihw.gov.au/reports/health-care-quality-performance/national-health-priority-areas-first-report/related-material [verified 16 May 2018].

15 Population Health Research Network. For researchers – data collections available. 2018. Available at: http://www.phrn.org.au/for-researchers/data-collections-available-by-jurisdiction/ [verified 21 June 2018].

16 Australian Institute of Health and Welfare. National health priority areas as at September 2017. 2017. Available at: https://www.aihw.gov.au/getmedia/28c917f3-cb00-44dd-ba86-c13e764dea6b/education-resource-health-priority-areas.pdf.aspx [verified 16 May 2018].

17 National Health and Medical Research Council. Research funding statistics and data. 2018. Available at: https://www.nhmrc.gov.au/grants-funding/research-funding-statistics-and-data [verified 16 May 2018].

18 de Lusignan S, van Weel C. The use of routinely collected computer data for research in primary care: opportunities and challenges. *Fam Pract* 2006; 23: 253–63. doi:10.1093/fampra/cmi106

19 National Statistical Service. A guide for data integration projects involving commonwealth data for statistical and research purposes. 2014. Available at: https://statistical-data-integration.govspace.gov.au/ [verified 16 May 2018].

20 Western Australian Data Linkage Branch. Data linkage WA. 2017. Available at: https://www.datalinkage-wa.org.au/sites/default/files/Dataset_Menu_Current_20170717.pdf [verified 16 May 2018].

21 SA–NT DataLink. Available datasets. 2018. Available at: https://www.santdatalink.org.au/available_datasets [verified 16 May 2018].

22 The Centre for Health Record Linkage. Master linkage key (MLK). 2018. Available at: http://www.cherel.org.au/master-linkage-key [verified 16 May 2018].

23 Stanley F, Glauert R, McKenzie A, O'Donnell M.. Can joined-up data lead to joined-up thinking? The Western Australian Developmental Pathways Project. *Healthcare Policy* 2011; 6(Spec Issue): 63–73.

24 Australian Government Productivity Commission. Data availability and use: overview & recommendations. Report No. 82. Canberra: Australian Government Productivity Commission; 2017.

25 Moore HC, Guiver T, Woollacott A, De Klerk N, Gidding HF. Establishing a process for conducting cross-jurisdictional record linkage in Australia. *Aust N Z J Public Health* 2016; 40: 159–64. doi:10.1111/1753-6405.12481

26 Population Health Research Network. Report on identifying and prioritising high value data collections for linkage. Population Health Research Network: Perth; 2017.

27 Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, Sørensen HT, von Elm E, Langan SM. The reporting of studies conducted using observational routinely-collected health data (RECORD) statement. *PLoS Med* 2015; 12: e1001885. doi:10.1371/journal.pmed.1001885

28 Population Health Research Network. Population Health Research Network strategic plan 2017–2026. Population Health Research Network: Perth; 2016.