Australian Clinical Trials Alliance

PHRN Population Health Research Network

USING LINKED ADMINISTRATIVE DATA IN CLINICAL TRIALS:

A Guide for Clinical Trialists and Researchers

Authors

Dr Felicity Flack
Dr Anna Kemp-Casey
Dr Natalie Wray

Acknowledgements

This guide has been produced to accompany the series of webinars hosted by ACTA in conjunction with the Population Health Research Network (PHRN) in March and April 2019. We gratefully acknowledge the review and guidance from the wider Australian Clinical Trials Alliance *Innovative Outcome Data* Reference Group, the support of the Australian Research Data Commons (ARDC) in providing the webinar platform, and the input from over 250 participants who attended the sessions. Recordings of these webinars are available online, please see the ACTA website for further information.

Please note that this guidance was accurate at the time of publication, readers should refer to the links provided for any updates.

ACTA gratefully acknowledges operational funding from the Australian Government's Medical Research Future Fund.
The webinar series and development of this guide were supported by the Population Health Research Network which is an initiative of the Australian Government being conducted as part of the National Collaborative Research Infrastructure Strategy.

Publication Details

Publication title: Using Linked Administrative Data in Clinical Trials: A Guide for Clinical Trialists and Researchers
Published: 2019
Publisher: Australian Clinical Trials Alliance
Online version: [insert link]
Suggested citation: Flack F, Kemp-Casey A and Wray N (2019). Using Linked Administrative Data in Clinical Trials: A Guide for Trialists and Researchers. Retrieved from the ACTA website www.clinicaltrialsalliance.org.au

CONTENTS

CHAPTER 1: DESIGNING CLINICAL TRIALS USING LINKED ADMINISTRATIVE DATA

Dr Felicity Flack

Introduction

Researchers collect a wide range of data about each participant during the course of a clinical trial. Linking this clinical trials data to administrative data can be useful in the context of clinical trials in a number of ways, including:

- The use of existing data, particularly from administrative sources
- A less burdensome method of data collection for participants during a clinical trial
- Ideal and cost-effective medium for long term follow-up of clinical trial participants
- Measurement of healthcare use before and after an intervention to assist in effectiveness and health economic analyses
- Post-market surveillance of therapeutics

The inclusion of linked administrative data in clinical trials in Australia is not a common design feature. In this chapter researchers will find advice about how to go about including linked data in their clinical trial design.

Objectives

By the end of this chapter you should be able to:

- Name examples of administrative data collections available for linkage to clinical trials cohorts
- Identify the benefits and limitations of using linked administrative data in clinical trials
- Describe the main factors to take into account when designing a clinical trial using linked administrative data

Administrative Data overview

Administrative data is the term given to those data collections that are made up of information that is routinely collected during the delivery of a service. (1) These data are collected via departments and agencies and used for policy, planning, management, monitoring, evaluation and research purposes.

The collection of administrative data is generally required or authorised by law. Typically, mandatory reporting requirements allow this data to be collected from health care providers without patient consent. (2) Administrative datasets usually contain both identifying information (name, address,

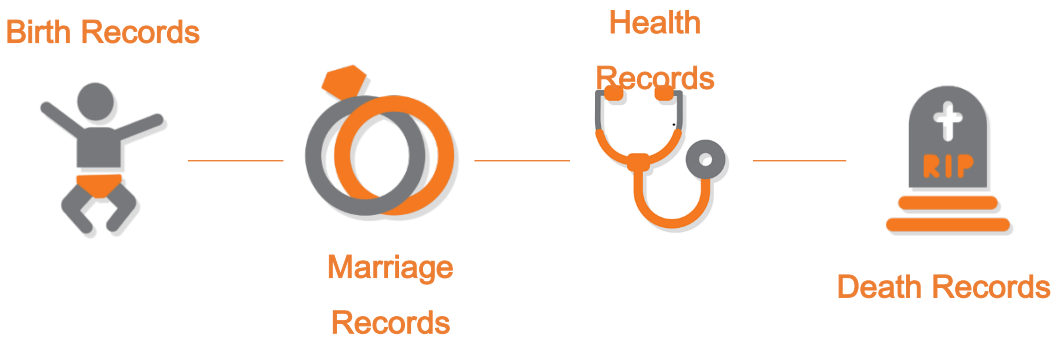
date of birth) and content information relating to individuals. It is important to note that the content information recorded in administrative datasets is only summary information (e.g. diagnosis, date of admission and treatment) as opposed to the much more detailed patient records used in the provision of patient care. (2)

The information is stored in secure data collections within each agency and all access to the data is controlled by a data custodian (an authorised nominee of the agency who holds the data).

Key features of administrative data

- Collected during service delivery
- Without consent
- Collection mandated by statute
- Used by government agencies for planning, monitoring and funding
- Population level coverage
- Availability of historical data
- Diversity of collections available

Figure 1: Administrative data across the lifespan



Who collects administrative data?

Australia is a federation and the responsibility for funding and running the health system is shared between the Australian Government and the state and territory governments. In addition, Australia has a substantial private health system. This federated health system means that some data is collected by the Australian Government, some by the state and territory governments and some by private sector organisations. Therefore, depending on the research question, using administrative data in Australia often requires linkage of data from these different sources.

Table 1:Examples of administrative data collections available for linkage in Australia

State/Territory Data Collections
• Admitted patients • Ambulance • Emergency Department • Mental Health • Disease-specific, procedural, device registries e.g. cancer registry • Perinatal • Birth Registrations • Death Registrations
Australian Government Data Collections
• Australian Immunisation Register • Australian Cancer Database • Medicare Benefits Schedule • Pharmaceutical Benefits Scheme • National Death Index

These data sources are available for linkage in most states and territories. Many more data sources are available for linkage e.g. pathology in individual jurisdictions. For more information about data sources available follow the links in the "Resources Available" section of this chapter.

There are few, large population level collections of data from private health organisations. However, some government data collections include data from the private sector e.g. admitted patient and perinatal data. The My Health Record is a large administrative data collection which includes data from private health services. *The Framework to Guide the Secondary Use of My Health Record System Data* has been approved. (3) Work towards implementing the Framework is in the early stages. The application process is unlikely to be clear before 2020.

Benefits of using linked administrative data in clinical trials

Administrative data is routinely collected, and its collection is funded by the relevant government agency. It is 'real world' data collected from the whole population over many decades. These data collections are often coded using internationally accepted coding systems e.g. ICD 10 AM. Therefore, administrative data has the potential to address some of the challenges with more traditional approaches to data collection for clinical trials including:

- The cost of data collection particularly for large numbers of participants
- Long term follow-up over many years
- Loss to follow-up

- Reporting and recall bias
- Health economic analysis (4, 5)

There are a range of ways that administrative data can be used to replace or augment traditional approaches to clinical trials data collection. (6)

- Long term follow-up (all participants or only those lost to follow-up)
- Pre-recruitment hospitalisation as a measure of trajectory to care
- Measurement of healthcare and/or pharmaceutical use before and after an intervention
- Measurement of primary and secondary endpoints
- Identification of eligible individuals or communities
- Economic analysis
- Disease prevalence, comorbidities, ethnicity and race
- Pragmatic or implementation clinical trials

Table 2: Examples of using linked administrative data in clinical trials

Post market Surveillance	
Whitstock MT, Pearce CM, Ridout SC, Eckermann EJ. A retrospective analysis of VIOXX in Australia: using clinical trial data and linked administrative health data to predict patient groups at risk of an adverse drug event (7)	Australian
Colvin L et al. Early morbidity and mortality following in utero exposure to selective serotonin reuptake inhibitors: a population-based study in Western Australia (8)	Australian
Kerr SJ et al All-cause mortality of elderly Australian veterans using COX-2 selective or non-selective NSAIDs: a longitudinal study (9)	Australian
Colvin L et al Off-label use of ondansetron in pregnancy in Western Australia (10)	Australian
Pearce A et al. Can administrative data be used to measure chemotherapy side effects? (11)	Australian
Health Economic Analysis	
Ward RL et al. Cost of cancer care for patients undergoing chemotherapy: The Elements of Cancer Care study (12)	Australian
Riley GF. Administrative and claims records as sources of health care cost data (13)	International
Gorham G, Howard K, Togni S, Lawton P, Hughes J, Majoni SW, et al. Economic and quality of care evaluation of dialysis service models in remote Australia: Protocol for a mixed methods study (14)	Australian
Watts CG, Cust AE, Menzies SW, Mann GJ, Morton RL. Cost-Effectiveness of Skin Surveillance Through a Specialized Clinic for Patients at High Risk of Melanoma. Journal of Clinical Oncology (15)	Australian

Cynthia Papendick et al. A randomized trial of a 1-hour troponin T protocol in suspected acute coronary syndromes: Design of the Rapid Assessment of Possible ACS In the Emergency Department with high sensitivity Troponin T (RAPID-TnT) study (16)	Australian
Frobert Thrombus Aspiration during ST-Segment Elevation Myocardial Infarction (17)	International
Taylor C et al. Hydroxyethyl starch versus saline for resuscitation of patients in intensive care: long-term outcomes and cost-effectiveness analysis of a cohort from CHEST (18)	Australian
Long Term Follow-up	
Hague WE et al. Long-Term Effectiveness and Safety of Pravastatin in Patients With Coronary Heart Disease: Sixteen Years of Follow-Up of the LIPID Study (19)	Australian
Dennis M, Kotchetkova I, Cordina R, Celermajer DS. Long-Term Follow-up of Adults Following the Atrial Switch Operation for Transposition of the Great Arteries - A Contemporary Cohort (20)	Australia
Gallagher M, Cass A, Bellomo R, Finfer S, Gattas D, Lee J, et al. Long-term survival and dialysis dependency following acute kidney injury in intensive care: extended follow-up of a randomized controlled trial (21)	International
d'Udekem Y, Iyengar AJ, Galati JC, Forsdick V, Weintraub RG, Wheaton GR, et al. Redefining expectations of long-term survival after the Fontan procedure: twenty-five years of follow-up from the entire population of Australia and New Zealand (22)	International

Limitations of using linked administrative data in clinical trials

Whilst there are a range of benefits to using administrative data in clinical trials it is extremely important to understand the limitations when designing a trial. (1, 4, 5) Administrative data is collected for administrative and financial purposes not for conducting research. The level of detail available may not be sufficient for all research purposes. When designing a clinical trial involving the use of administrative data, researchers should consider the following:

Limited data items – administrative data collections include a limited set of data items related to the purpose of the data collection. Therefore, they often do not include data on a range of confounders and risk factors such as height or weight. In general, they permit crude risk adjustment; e.g., individual morbidities, Charlson Comorbidity Index, but lack sufficient clinical detail to risk adjust for contemporary algorithms.

Uncertain validity – How good administrative data is at measuring the desired exposures or outcomes (validity) can vary between data collections and different data items. For example, changes in coding directives and practices over time can affect validity. (23) This is particularly true for collections where their use for research purposes is relatively new. The validity will depend on the correct information being available, accurate coding and accurate data entry. Sometimes data custodians provide information about the validity of the data. Researchers should conduct their own validity studies. (24, 25)

Limited metadata – In order to design a clinical trial using administrative data, researchers require information about the scope or purpose of the data collection, the data items available, the years covered and any changes in coding or policy over time that may affect the analysis. Without this information there is the potential for misleading interpretation of the analysis. In the absence of detailed metadata researchers should seek advice from experienced clinical coders or other researchers with experience using the data.

Linkage quality/linkage error – linkage quality can be affected by many factors including:

- Quality of the data to be linked
- The quantity of data items linked i.e. the more data items used in the linkage process the higher the quality of linkage e.g. full name, address, sex and date of birth will give a better result than sex and date of birth.
- The linkage methodology e.g. deterministic, probabilistic
- The quality assurance processes

Key points to consider

Is the data I need available?

First decide whether you need data from a single jurisdiction or multiple jurisdictions. In many clinical trials access to data from multiple jurisdictions will be essential e.g. if death is the primary outcome then this information is required even if the death occurs outside the state in which the trial is being conducted.

Single Jurisdiction Example

If you are only interested in if the trial participants have been hospitalised in the state or territory in which the clinical trial is being conducted. This would require data from a single jurisdiction e.g. NSW.

Start your search for data availability on the website of the state or territory data linkage unit.

http://www.cherel.org.au/master-linkage-key

Check whether hospital data (admitted patient data) is on the list of data collections routinely linked.

If it is routinely linked or is on the list of data collections that have been linked in the past then it will be available for linkage to your clinical trials cohort.

Now check to see if there is more information about the data collection including:

- The purpose and history of the data collection
- The years the data is available
- The data items (variables available)
- Any changes in the collection or data items over time
- What coding system is used and if any coding directives are announced from time to time, which may distort the counts for the variable effected
- Whether it includes data from both the public and private health systems
- Is there a different lag period for rural and remote sites compared to cities

http://www.cherel.org.au/data-dictionaries#section1

Use this information to determine whether the data will be suitable for use in your clinical trial. If you have any questions about the data contact the data linkage unit and/or the data custodian. The client services team at the data linkage unit is usually the best first point of contact.

Cross-Jurisdiction Example

You may want linked data from multiple jurisdictions. For example, you may be interested in if the participants in a clinical trial have been hospitalised in any state or territory and also whether they have died. This would require admitted patient data from all states and territories and the National Death Index.

Start your search for data availability on the Population Health Research Network website.

https://www.phrn.org.au/for-researchers/data-collections-available/

The information on this website will enable you to compare availability across jurisdictions.

For each jurisdiction you will need to review the information available and determine whether the data will be suitable for your clinical trial.

When will the administrative data be available?

It is also important to find out when the data you require will be available. Most administrative data are not collected and linked in real time. There will be a lag between when the event of interest occurred and when this data is available to be linked and accessed for analysis. The lag can be as long as 12 months.

Is the administrative data fit for purpose?

Once the availability of the data has been confirmed the researcher should consider whether the data will be able to deliver scientifically valid results.

One of the important considerations is what codes and algorithms will be used to classify exposures, outcomes, confounders, and effect modifiers.

Some data linkage units and data custodians provide validation data e.g. www.cherel.org.au/validation-studies

It is also worth conducting a literature search for studies which have previously used the data collection of interest. There may be validation studies or some validation information in these publications.

If you are designing a trial which will link administrative data from multiple jurisdictions, such as hospital data from all states and territories, you will need to check whether the variables you require are available from each jurisdiction. In addition, check that the same level of information is available. For example, some jurisdictions may release full date of hospital admission and others may not.

The exact linkage methodology will have to be determined. The nature and quality of the linkage will need to be taken into consideration in the analysis of the data. The linkage methodology will be determined by whether it is a single site, multi-site, single jurisdiction or cross-jurisdiction clinical trial. Sometimes the data collections to be linked will also impact on the linkage methodology. The data linkage units are the experts in linkage methodology and will work closely with the research team to develop a feasible approach for each clinical trial.

This is also a good time to consider what information will be required for publication of the results. The RECORD Checklist is a useful tool to assist researchers to ensure they have all the information required for publication when they have used administrative data.

See www.record-statement.org/checklist.php or for the pharmacoepidemiology version www.record-statement.org/checklist-pe.php .

What skills/experience will the research team require to analyse administrative data?

As described above, administrative data has a number of features that make it different from data collected specifically for clinical trials. Therefore, it requires considerable cleaning and preparation prior to analysis and different approaches to analysis. Ideally the research team will include a biostatistician or analyst with experience in the analysis of administrative data. If they have experience with the specific data collections used, the analysis will likely run more smoothly. If you do not have access to an experienced analyst, it is highly recommended that at least one team member attends one of the analysis of linked data courses available. See the Resources section for more information.

How will consent be managed?

In the design stage of your clinical trial be sure to ask the participants for consent to link to administrative data. However, if your trial has already commenced or been completed it may be possible to have the requirement for consent to link administrative data waived. The ethical issues around consent and waivers of consent will be covered in Chapter 3 of this guide.

When does the data need to be linked?

You will need to decide when your data should be linked over the life of your clinical trial. Depending on the reason linked data has been included in the design the linkage could occur once after the completion of enrolment or trial completion. This is the simplest approach. However, it may be necessary for linkage to occur multiple times throughout the trial e.g. at 50% enrolment, at 100% enrolment, at end of follow-up, 5 years later.

How will the data be managed?

All clinical trials data must be managed carefully and securely. All data custodians and ethics committees will require you to submit a detailed data management plan. The Australian National Data Service (now ARDC) has a range of resources to assist you to develop an appropriate data management plan (www.ands.org.au/working-with-data/data-management/data-management-plans).

Involving administrative data in your clinical trial may add to the complexity of the data flows and may result in additional data management requirements. In particular the data linkage variables will need to be separated from the content data and transferred to the data linkage unit. The administrative data (content) will be transferred from the data custodian to a suitable facility for cleaning and analysis. This may be at the researcher's home institution or it may be in a secure facility designated by the data custodians. Some data custodians and/or ethics committees may require that only some members of the research team have access to the linked data.

As part of the data management plan it can be helpful to create a data flow diagram which describes where data is collected, by whom and where it is transferred and stored throughout the clinical trial. This will help you assess the level of risk of identification at different stages of the trial and therefore what the appropriate level of information security should be. You will need to work with your data linkage unit/s to describe the data flow for the linkage part of the trial.

Real examples of linkage diagrams can be found in the literature. (14, 26)

Resources available

Metadata

Population Health Research Network

www.phrn.org.au/for-researchers/introduction/

Australian Data Linkage Units

www.phrn.org.au/about-us/who-is-involved/australian-data-linkage-units/

Data Management Frameworks and Plans

www.ands.org.au/working-with-data/data-management

Training courses

University of Western Australia, Perth

Introductory Analysis of Linked Health Data
http://handbooks.uwa.edu.au/unitdetails?code=PUBH5785

Advanced Analysis of Linked Health Data
http://handbooks.uwa.edu.au/unitdetails?code=PUBH5802

University of Sydney, Sydney

Introductory Analysis of Linked Data (PUBH5215) - Professional Development Course
www.sydney.edu.au/courses/units-of-study/2019/pubh/pubh5215.html

WHO Collaborating Centre for Viral Hepatitis, Melbourne

Advanced Analysis of Linked Health Data
www.doherty.edu.au/news-events/events/advanced-analysis-of-linked-health-data-course

References

1. Davies J, Barnes H, Dibben C. Education Administrative Data: Exploring the Potential for Academic Research. St Andrews, Scotland: Administrative Data Liaison Service; 2010.
2. Allen J, Holman CD, Meslin EM, Stanley F. Privacy Protectionism and Health Information: Is There Any Redress for Harms to Health? Journal of Law & Medicine. 2013;21(2):473-85.
3. The Framework to Guide the Secondary Use of My Health Record System Data. Department of Health, Commonwealth of Australia; 2018.
4. Jorm L. Routinely Collected Data as a Strategic Resource for Research: Priorities for Methods and Workforce. Public Health Research & Practice. 2015;25(4):30.
5. Hashimoto RE, Brodt ED, Skelly AC, Dettori JR. Administrative database studies: goldmine or goose chase? Evidence-based spine-care journal. 2014;5(2):74-6.
6. Makady A, Goettsch W. GetReal - Project No. 115546 WP1: Deliverable D1.2 Review of current policies/perspectives. Zorginstituut Nederland.
7. Whitstock MT, Pearce CM, Ridout SC, Eckermann EJ. A retrospective analysis of VIOXX in Australia: using clinical trial data and linked administrative health data to predict patient groups at risk of an adverse drug event. Aust N Z J Public Health. 2010;34(4):431-2.
8. Colvin L, Slack-Smith L, Stanley FJ, Bower C. Early morbidity and mortality following in utero exposure to selective serotonin reuptake inhibitors: a population-based study in Western Australia. CNS Drugs. 2012;26(7):e1-14.
9. Kerr SJ, Rowett DS, Sayer GP, Whicker SD, Saltman DC, Mant A. All-cause mortality of elderly Australian veterans using COX-2 selective or non-selective NSAIDs: a longitudinal study. Br J Clin Pharmacol. 2011;71(6):936-42.
10. Colvin L, Gill AW, Slack-Smith L, Stanley FJ, Bower C. Off-label use of ondansetron in pregnancy in Western Australia. Biomed Res Int. 2013;2013:909860.
11. Pearce A, Haas M, Viney R, Haywood P, Pearson SA, van Gool K, et al. Can administrative data be used to measure chemotherapy side effects? Expert Rev Pharmacoecon Outcomes Res. 2015;15(2):215-22.
12. Ward RL, Laaksonen MA, van Gool K, Pearson SA, Daniels B, Bastick P, et al. Cost of cancer care for patients undergoing chemotherapy: The Elements of Cancer Care study. Asia Pac J Clin Oncol. 2015;11(2):178-86.
13. Riley GF. Administrative and claims records as sources of health care cost data. Med Care. 2009;47(7 Suppl 1):S51-5.
14. Gorham G, Howard K, Togni S, Lawton P, Hughes J, Majoni SW, et al. Economic and quality of care evaluation of dialysis service models in remote Australia: protocol for a mixed methods study. BMC Health Serv Res. 2017;17(1):320.
15. Watts CG, Cust AE, Menzies SW, Mann GJ, Morton RL. Cost-Effectiveness of Skin Surveillance Through a Specialized Clinic for Patients at High Risk of Melanoma. J Clin Oncol. 2017;35(1):63-71.
16. Papendick C, Blyth A, Seshadri A, Edmonds MJR, Briffa T, Cullen L, et al. A randomized trial of a 1-hour troponin T protocol in suspected acute coronary syndromes: Design of the Rapid Assessment of Possible ACS In the emergency Department with high sensitivity Troponin T (RAPID-TnT) study. Am Heart J. 2017;190:25-33.
17. Fröbert O, Lagerqvist B, Olivecrona GK, Omerovic E, Gudnason T, Maeng M, et al. Thrombus Aspiration during ST-Segment Elevation Myocardial Infarction. New England Journal of Medicine. 2013;369(17):1587-97.
18. Taylor C, Thompson K, Finfer S, Higgins A, Jan S, Li Q, et al. Hydroxyethyl starch versus saline for resuscitation of patients in intensive care: long-term outcomes and cost-effectiveness analysis of a cohort from CHEST. Lancet Respir Med. 2016;4(10):818-25.
19. Hague WE, Simes J, Kirby A, Keech AC, White HD, Hunt D, et al. Long-Term Effectiveness and Safety of Pravastatin in Patients With Coronary Heart Disease: Sixteen Years of Follow-Up of the LIPID Study. Circulation. 2016;133(19):1851-60.
20. Dennis M, Kotchetkova I, Cordina R, Celermajer DS. Long-Term Follow-up of Adults Following the Atrial Switch Operation for Transposition of the Great Arteries - A Contemporary Cohort. Heart Lung Circ. 2018;27(8):1011-7.
21. Gallagher M, Cass A, Bellomo R, Finfer S, Gattas D, Lee J, et al. Long-term survival and dialysis dependency following acute kidney injury in intensive care: extended follow-up of a randomized controlled trial. PLoS Med. 2014;11(2):e1001601.
22. d'Udekem Y, Iyengar AJ, Galati JC, Forsdick V, Weintraub RG, Wheaton GR, et al. Redefining expectations of long-term survival after the Fontan procedure: twenty-five years of follow-up from the entire population of Australia and New Zealand. Circulation. 2014;130(11 Suppl 1):S32-8.
23. Knight L, Halech R, Martin C, Mortimer L. Impact of changes in diabetes coding on Queensland hospital principal diagnosis morbidity data. Health Statistics Centre, Queensland Health; 2011. Contract No.: Technical Report #9.
24. Seeskin ZH, Ugarte G, Datta AR. Constructing a toolkit to evaluate quality of state and local administrative data International Journal of Population Data Science. 2019;4(1).
25. Tran DT, Jorm L, Lujic S, Bambrick H, Johnson M. Country of birth recording in Australian hospital morbidity data: accuracy and predictors. Australian and New Zealand Journal of Public Health. 2012;36(4):310-6.
26. 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(2):159-64.

CHAPTER 2: ACCESSING LINKED DATA FOR CLINICAL TRIALS

Dr Felicity Flack and Dr Natalie Wray

Objectives

By the end of this chapter you should be able to:

- Explain the three key steps in the data linkage process (separation, linkage and access)
- Name some of the data linkage services and facilities available
- Describe the 6 steps in the process to access linked administrative data
- Understand the current timelines and costs involved in accessing linked administrative data

Data Linkage overview

Data linkage, in simple terms, is a method of bringing together information derived from different sources but relating to the same individual or event into a single file. (1) Data linkage enables the linkage of information about people, places and events in a way that protects individual information privacy. A range of methods and technology are used to link existing information from different data collections which can then be provided in a privacy-protected format to approved researchers for a range of population-based studies. Data linkage relies on the availability of shared unique identifiers across data collections. Common identifiers used to link data include: name, address, sex, date of birth and record date. (2)

There are three distinct groups involved in the data linkage and access process. (3)

Data Custodians

Data custodians are the people who look after the data collections. They work within an organisation or agency (such as a government health department) and are responsible for the secure collection, use and disclosure of data. Data custodians collect and store identifying information (e.g. name, sex, date of birth) and also the related content information (e.g. health information such as diagnosis and treatment details).

Data Linkers

Data linkers are the people who use identifying information to create and maintain linkage IDs which allow data to be linked within and between data collections. Data linkers usually work in a data linkage unit that is either within or associated with a government agency.

Researchers

Researchers are the people who use the data for the purpose of analysis and research. This process is only possible after approval by a Human Research Ethics Committee (HREC) and all relevant data custodians.

How is Data Linked for Research?

The following is a broad description of the overall approach to linking data in Australian data linkage units. This approach is sometimes called “the separation principle”. (3) It is designed to minimise the privacy risks to research participants through the separation of identifiers from content data and also through the separation of the roles and responsibilities of the staff with access to data. The specific details of how data is linked for research will vary from unit to unit. Methods used internationally may differ from the description below. (4)

Separation and Provision of Identifiers to Data Linkage Units

The data custodian makes a copy of the identifiers (sometimes called linkage variables) from each record (e.g. name, address, sex, date of birth) in their data collection and sends it together with a record ID to the data linkage unit. For routinely linked data sources, data custodians provide regular updates of the linkage variables and record IDs to the data linkage unit.

These records do not include information relating to the health of an individual or other content information about them other than the database of origin. The portion of the record containing the content information (e.g. health information such as diagnosis and treatment details) remains with the data custodian, meaning that the data linkers never have access to content data. This separation of identifiers and content information is an important step in protecting the privacy of information about individuals. It ensures that the data custodians remain the only people that have access to fully identified information about a person.

Creation of Linkage IDs

To allow data about the same person to be linked within data collections and across different data collections, data linkers within a data linkage unit create unique linkage IDs. Upon receiving linkage variables and record IDs for individuals from data custodians, data linkers check to see whether each person has an existing linkage ID using a statistical *probability method*. If a person already has a linkage ID, the new record ID is assigned to their linkage ID. If a person is judged to be new to the system, the data linkers create a new linkage ID which is then assigned to that person. The linkage IDs are stored on secure computer servers and can only be accessed by authorised data linkage unit staff (i.e. data linkers).

Provision of linked data to researchers

If the researcher’s project is approved by a Human Research Ethics Committee (HREC), all the relevant data custodians and is deemed feasible by the data linkage unit, then the data linkers use the Linkage IDs to create Project Linkage IDs that are unique for each individual and specific for the approved study. They then send the Project Linkage IDs along with the Record IDs of the required records to the data custodians. Using the Record IDs, the data custodians extract approved data from the specified records from their collections and place them in a new file. The data custodian then replaces the personal information of each record with its matched Project Linkage ID. The researcher is provided with the content data of each record and its corresponding Project Linkage ID by each data custodian.

The researcher uses the Project Linkage ID to determine which records from different datasets belong to the same person, without having access to the personal information, in order to create a merged dataset for their analysis.

Data linkage methods

There are two main methods, reported in the research literature, used to link data. These are probabilistic and deterministic.

Probabilistic linking is the linkage of records from two (or more) files and is based on the probabilities of agreement and disagreement between a range of match variables. (5)

Deterministic linking is the matching of two records based on exact agreement of the selected unique identifiers. This method requires a unique or near unique identifier and high-quality data files. (6, 7)

TABLE 1: DATA LINKAGE METHODS

Method	Uses and Advantages	Disadvantages
Probabilistic linking systems	• Linkage quality (accuracy) can be refined using individual parameters/fields • Highly adaptable – easily accommodates a growing number of fields in data files • Based on proven statistical underpinnings • Reflects the uncertainty that exist for potential links • Once implemented, does not require much manual tuning to optimise and maintain • Works best when used in conjunction with deterministic methods	• Argued to be more complex and less transparent than other approaches • Implementation overhead • Required specialised skills, limited available expertise in probabilistic methods
Deterministic linking systems	• Uses straight forward transparent rules to make links • Easy to implement and maintain • Clerical review not usually required • Used when files have high quality data with powerful discriminating linking keys	• Matches depend on error rate of the linking variables • Linkage quality not always easy to refine • Matching rules can become complex when there are many data fields

	• Can incorporate advanced computational methods and probabilistic components	

Cross-Jurisdictional Data Linkage

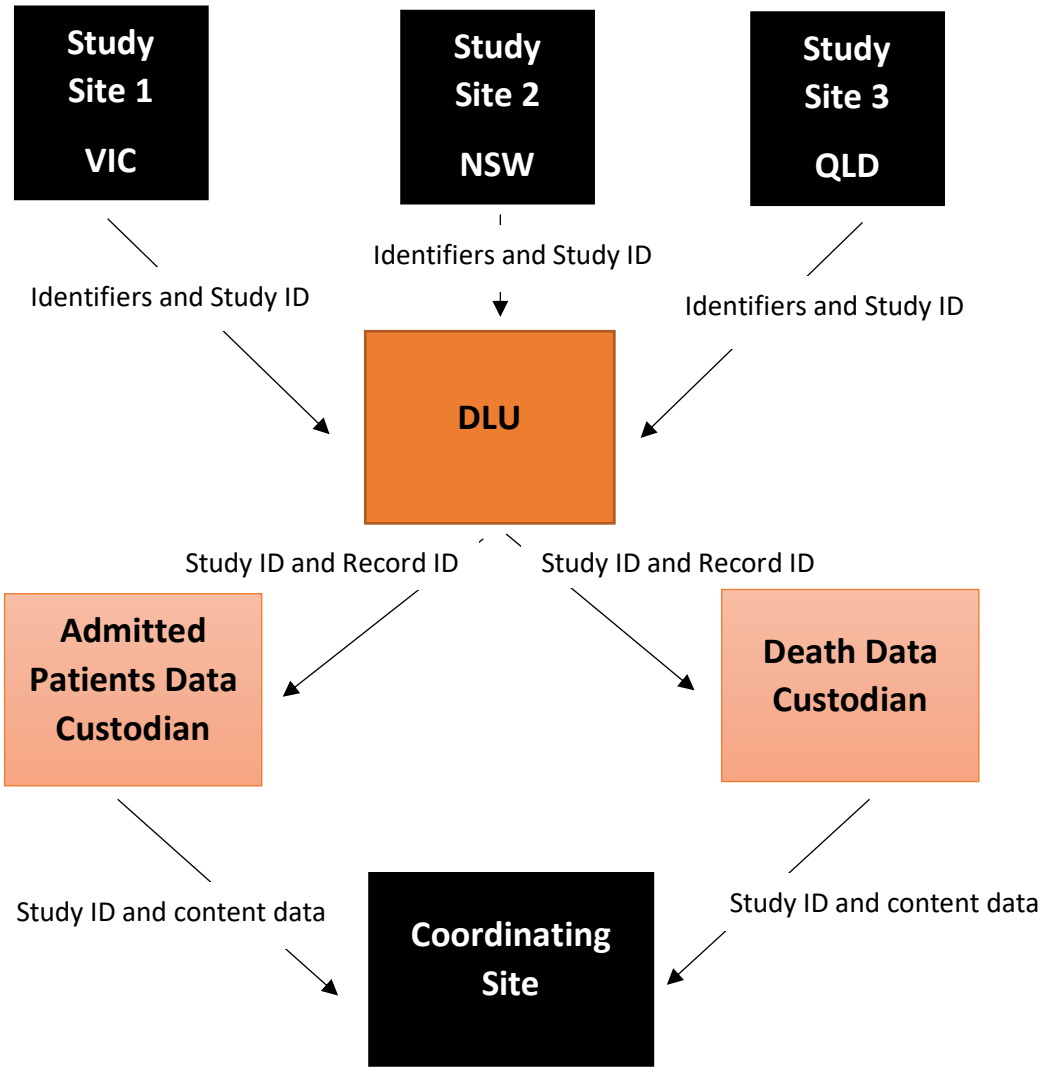
Sometimes there are benefits to linking data collections from different jurisdictions e.g. a clinical trial cohort to admitted patient data in every state and territory. This is called cross-jurisdictional data linkage.

The reasons that cross-jurisdictional data linkage may be used include:

- To increase the statistical power for research on rare conditions or outcomes e.g**.** data from more than one state or territory may be necessary to achieve the required sample size.
- One jurisdiction does not collect all the data required to answer the research question e.g. linking state hospital data to Commonwealth Pharmaceutical Benefits System data for post-market surveillance of a new drug.
- To assess cross-border usage of services.
- To obtain accurate data for longitudinal studies e.g. linkage to the national death register to determine if study participants have died. If a study participant dies outside the state in which the study is conducted this will not be registered on the state death registry.

FIGURE 1: DATA FLOW DIAGRAM

Example: How could my clinical trial cohort be linked to admitted patient data?



The Population Health Research Network (PHRN)

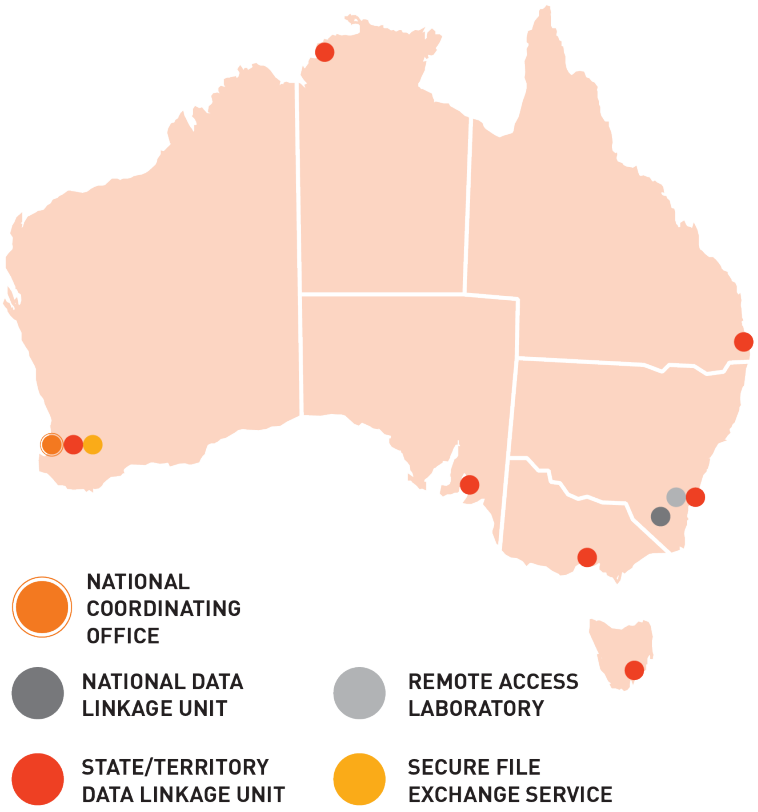
The PHRN is an initiative of the Australian Government's National Collaborative Research Infrastructure Strategy (NCRIS). It was established in 2009 to build a nationwide data linkage infrastructure capable of securely and safely managing health and related information from around Australia.

The PHRN does not collect or hold personal health information. The PHRN's role is simply to support units around Australia that link data and to enable researchers' access to linked information from existing data collections.

Who is involved?

The PHRN is not in itself an organisation, rather a network of collaborating organisations spread across Australia.

FIGURE 2: THE PHRN



PHRN Regional Data Linkage Services and Facilities

The PHRN collaboration involves six data linkage units which service each of the states and territories.

TABLE 2: PHRN REGIONAL DATA LINKAGE SERVICES AND FACILITIES

Data Linkage Unit	Location	Jurisdiction
WA Data Linkage Branch	Department of Health, WA	Western Australia
SA-NT DataLink	University of South Australia	SA and NT
CHeReL	Department of Health, NSW	NSW and ACT
Centre for Victorian Data Linkage	Department of Health, VIC	Victoria
Tasmanian Data Linkage Unit	Menzies Institute for Medical Research	Tasmania
Data Linkage Queensland	Queensland Health	Queensland

PHRN National Data Linkage Services and Facilities

National Data Linkage Unit

The AIHW as an accredited Commonwealth Integrating Authority offers national data linkage services.

PHRN Online Application System

The PHRN Online Application System provides a unified online form to apply for access to cross-jurisdictional linked data. The system will allow you to:

- Use the same form for requests for quotes, expression of interest and formal applications
- Use the one form to obtain approval from multiple jurisdictions
- Submit your application to multiple jurisdictions simultaneously
- Track the progress of your application for data
- Share the application with multiple researchers for viewing and comments; and
- Manage all relevant documents in a central point.

The online application system is available at https://oas.phrn.org.au

Secure Remote Access Facility

The Sax Institute have developed the Secure Unified Research Environment (SURE) as part of the PHRN. SURE is a remote-access computing environment that allows researchers to access and analyse linked health-related data files for approved studies in Australia.

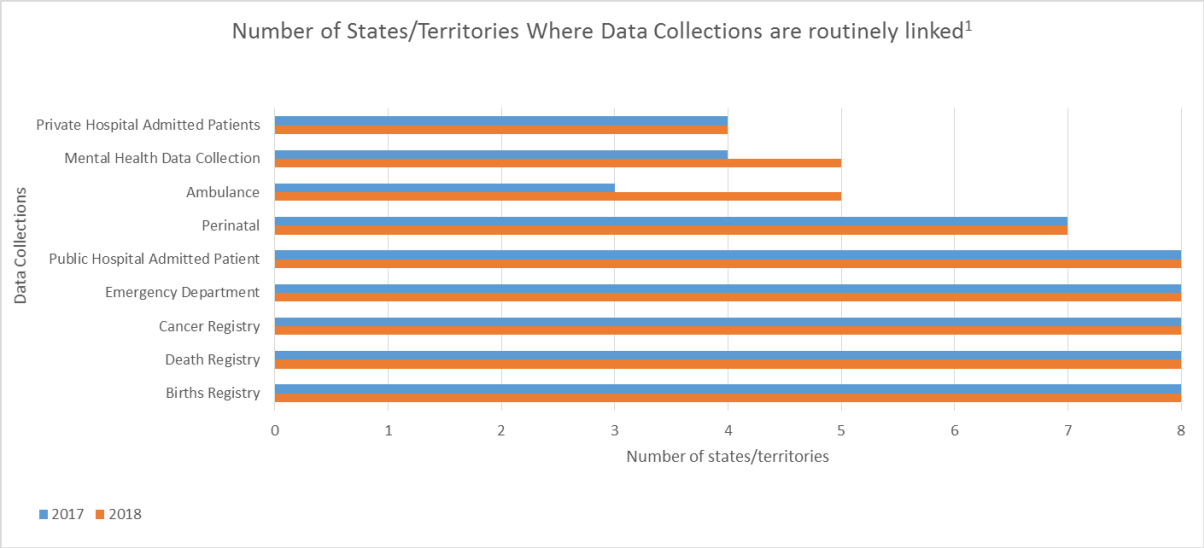
Secure File Transfer

Transferring data from one place to another is an important component of an appropriate data management plan. SUFEX is a secure file transfer service for the Population Health Research Network (PHRN) and its stakeholders. SUFEX is one of the options that data custodians can use to send files to the PHRN data linkage units and researchers. SUFEX uses a secure online application that allows users to send and receive files from anywhere at any time. It provides users with a secure file exchange service and is not a file storage solution

Metadata and other information available

For information about routinely linked data across Australia a good place to start is the PHRN website. Here you can find summary information about the data collections routinely linked and the comparative availability from different jurisdictions.

FIGURE 3: DATA COLLECTIONS ROUTINELY LINKED BY NUMBER OF JURISDICTIONS



For more detailed information you may need to look at the websites of individual data linkage units. See Table 2 for links.

Applying to Access Linked Data

Clinical Trial Design

The first step in the process for applying and accessing linked data is to design your trial using data and methods appropriate for achieving the aims of the trial. See Chapter 1 for more information about designing clinical trials using linked data.

Feasibility

Once you have designed your clinical trial you will need to complete a data application. The data application describes your trial, the data to be collected, linked and used and how you will manage the data to ensure privacy and confidentiality risks are minimised.

If you are including linked data from a single jurisdiction, e.g. one state, you will complete the data application form required by that jurisdiction. If you are including linked data from multiple jurisdictions, then you should complete the online form in the PHRN online application system.

Each data linkage unit has a client services team. The client services teams can assist researchers with designing clinical trials. In particular with:

- Information about data collections
- Information about linkage methods and requirement
- Information and advice about approval requirements
- Negotiation with data custodians for access to data
- Advice about feasibility
- Provision of quotes and cost estimates

When you submit your data application the client services team will review your application and either request more information or advise you that it is ready for submission for data custodian approval and/or ethics approval.

Data Custodian Approval

For access to administrative data there is usually a legislative requirement for each data custodian to give approval for access.

Ethics Approval

Approval from at least one HREC is required for clinical trials involving linked data. However, if more than one jurisdiction is involved, more than one HREC approval may be required. Linked data projects are currently not included in the National Mutual Acceptance Scheme in several states. In some jurisdictions, approval by the state health department HREC is mandatory.

More information about ethics requirements is provided in Chapter 3.

In some jurisdictions, a research governance review is also required. The governance review is a separate process to the ethical and scientific review. Data linkage units can advise on local requirements for research governance review.

Linkage

Once all approvals have been received, the relevant linkage unit/s can commence the requested linkage. Please note that most data linkage units operate over capacity and there is usually a queue for linkage.

If you have requested data to be linked to a clinical trials cohort you will need to provide the linkage variables to the data linkage unit/s in the format specified by the data linkage unit. You should already have agreed which linkage variables and the format during the design phase.

Example of linkage variables

Data format	Character-separated text (e.g. .csv, .tab), fixed-width text or Excel spreadsheet
Data fields supplied for linkage ***Every record must have a unique ID***	Project record ID, First name, Middle name, Surname, Address, Suburb, Postcode, State, Date of birth (day), Date of birth (month), Date of birth (year), Sex

You will also have decided who in the research team will be responsible for separating the linkage variables and transferring them to the linkage unit/s.

Data Extraction and Delivery

When the data linkage unit/s have completed the linkage and created the project specific linkage keys they will contact the relevant data custodians to extract the data for your clinical trial. The data custodians will extract the data and attach the relevant project specific linkage keys. The content data will then be transferred to the agreed data access facility.

When you gain access to the linked data the first task is to check that the data delivered matches the agreed specifications and approvals. If not, you may need to contact the data linkage unit and request a re-extraction.

Once you have confirmed receipt of the correct data you will need to clean and format the data prior to commencing analysis.

Checking, cleaning and formatting data is a significant part of the research process so please allocate adequate time i.e. 3-6 months. See Tran DT et al for an example of the steps required for checking and cleaning (9).

Approval Times

Each data linkage project is bespoke. The linkage and data provided is tailored specifically for the needs of each particular project. Therefore, it is not possible to give definitive times and costs. However, there are a number of things that impact both time and cost:

Complexity – the more complex a project, the longer and more expensive it is likely to be. Things that increase complexity include:

- The requirement for new linkages
- The number of data collections
- Complex study designs and cohort selection e.g. case/control designs
- The number of jurisdictions
- The number of approvals required – the more approvals, the more time
- Researcher response times – the quicker researchers respond to questions from the data linkage unit, data custodian and/or HREC the quicker the application will be approved.

Costs

It is always a good idea to request a quote early on so you have an indication of the cost of your research proposal. Each data linkage unit have their own pricing and/or charging policy. There may be charges for client services, linkage, geocoding or extraction. The charges depend on the size and complexity of the linkage or extraction tasks involved. For extractions, the cost depends on the number of datasets, number of years and number of individuals in the data extract.

Charges also apply for linked data to be analysed and stored in the SURE (remote access laboratory) and to lodge an application to the AIHW HREC.

For cross-jurisdictional projects, researchers can submit a request for quote via the PHRN Online Application System.

For single-jurisdiction projects, researchers should discuss a cost estimate with the Client Services Team at the required state/territory data linkage unit.

References

1. Hobbs MS, McCall MG. Health Statistics and Record Linkage in Australia. Journal of Chronic Diseases. 1970;23(5):375-81.
2. Sibthorpe B, Kliewer E, Smith L. Record Linkage in Australian Epidemiological Research: Health Benefits, Privacy Safeguards and Future Potential. Aust J Public Health. 1995;19(3):250-6.
3. Population Health Research Network. Who is involved? Australia: Population Health Research Network; 2019. Available from: https://www.phrn.org.au/about-us/data-linkage/whos-involved/
4. Population Health Research Network. How is data linked? Australia: Population Health Research Network; 2019. Available from: https://www.phrn.org.au/about-us/data-linkage/how-is-data-linked/
5. Kelman CW, Bass AJ, Holman CD. Research Use of Linked Health Data-A Best Practice Protocol. Aust N Z J Public Health. 2002;26(3):251-5.
6. Karmel R, Anderson P, Gibson D, Peut A, Duckett S, Wells Y. Empirical Aspects of Record Linkage Across Multiple Data Sets Using Statistical Linkage Keys: The Experience of the PIAC Cohort Study. BMC Health Serv Res. 2010;10:41.
7. Gill L. Methods for Automatic Record Matching and Linking and Their Use in National Statistics. Statistics N, editor. London2001.
8. Christen P, Goiser K. Quality and Complexity Measures for Data Linkage and Deduplication. In: Guillet F, Hamilton HJ, editors. Quality Measures in Data Mining. Studies in Computational Intelligence. 43. Berlin: Springer; 2007. p. 127-51.
9. Tran DT, Havard A, Jorm LR. Data cleaning and management protocols for linked perinatal research data: a good practice example from the Smoking MUMS (Maternal Use of Medications and Safety) Study. BMC medical research methodology. 2017;17(1):97-.

CHAPTER 3: ETHICS AND LINKED ADMINISTRATIVE DATA IN CLINICAL TRIALS

Dr Felicity Flack

Objectives

By the end of this chapter you should be able to:

- Explain the key ethical issues related to accessing linked administrative data
- Describe the legal framework applicable to accessing linked administrative data without consent
- List some strategies to demonstrate respect in the absence of consent

Core Ethical Values

The National Statement 2007 (updated 2018) describes four core values which underpin the ethical relationship between researchers and research participants: (1)

- research merit and integrity
- justice
- beneficence
- respect

Human Research Ethics Committees members are required to consider all four core values in deciding whether the research is justified. Clinical trials researchers will be familiar with the core values and accustomed to taking them into account when designing and conducting all of their research. In this section we consider how these values relate to the use of linked data in clinical trials.

Key Ethical Considerations

Key Ethical Issues

Research Merit and Integrity

The National Statement says that research is only ethically justifiable if it has merit and the researchers who are to carry out the research have integrity. HREC members must consider whether the project as a whole is worth doing. A particular issue HRECs will consider for projects involving linked data is whether the data available and methodology proposed can answer the question that is being proposed. They will also consider whether the researchers have the necessary training and

experience in merging and analysing linked data. See Chapter 1 for more information about designing clinical trials which use linked data.

Beneficence

The National Statement requires that "*the likely benefit of the research must justify the risks of harm or discomfort to the participants*." HRECs must assess "*the risks of harm and the potential benefits to the research participants and the wider community*." Additionally, the National Statement requires that researchers should be "sensitive to the welfare and the interests of people involved in their research and reflect on the social and cultural implications of their work".

The benefits to the participants of using linked data will vary from trial to trial. Benefits may include reducing the number of trial visits and the volume of data collected directly from the participants. Broader benefits may include cost savings and efficiencies in data collection, the inclusion of data collected prior to enrolment and longer term follow up data than may be feasible in a conventional clinical trial as well as the ability to undertake health economic analysis of interventions.

There may be some slightly increased data security risks related to data being collected over a longer period of time and transferred and stored in multiple locations. However, compared to the existing risks associated with the volume, detail and personal nature of clinical trials data the additional risks of using linked data are likely to be minimal.

Data Management

Careful data management is important to minimise and manage risks and to demonstrate respect to participants. All clinical trials should have a data management plan. For projects where participants have not provided consent to participate even more care should be taken with data management. There are four main steps in good data management:

1. **Identify Data Sources**
   Identify all the data that will be collected or created during the project. At a minimum write a list of this data but preferably draw a flow diagram which describes how the data will move from place to place over the course of the project.

2. **Identify who will have Access to What Data**
   Look at your flow diagram and determine who needs access to the data at the different stages of the project. Not all investigators will require access to the individual level linked data.

3. **Evaluate Risks**
   The severity of the risks associated with the use of linked data is mostly associated with the identifiability of the data. The risks of each step in your flow diagram should be evaluated.

   Identifiability is not an inherent quality of data and can be affected by:

   - The type of information e.g. full date of birth is more identifying than year of birth

- The quantity of information i.e. the more information you have about a person the easier it is to identify them
- The other information available to the person who has access to the data
- The skills and technology available to the person who has access to the data

The National Statement (NS 3.1 element 4) includes helpful information on evaluating risks of identifiability.

4. Develop a Data Management Plan

A data management plan should cover all stages of the research project including data collection, creation, access, use, analysis, disclosure, storage, retention, disposal, sharing and re-use. The National Statement (NS 3.1.5) says that a data management plan should include:

- Physical security
- Policies and procedures
- Contractual and licencing arrangements and confidentiality agreements
- Training for the research team
- The form in which data will be stored
- The purposes for which the data will be used and/or disclosed
- The conditions under which access to the data may be granted to others
- What information from the data management plan needs to be communicated to the participants.

There is a range of information available to assist you in developing a suitable data management plan for your clinical trial. The Australian Research Data Commons (ARDC) provides a guide to developing a data management plan which may be a helpful start. (2) The Ten Simple Rules for Creating a Good Data Management Plan is also a good guide. (3) Many universities have template data management plans that can be adapted for individual use.

Access with Consent

For many uses of linked data in clinical trials it will be possible to gain consent from the participants to link their data to other sources of information about them. This will necessitate a separate explicit statement in the trial consent form that needs to be signed and dated by the participant.

Key Ethical Issues

Respect

According to the National Statement, respect for human beings is recognition of their intrinsic value. (1) It means recognising the value of human autonomy - the capacity to make one's own decisions including decisions about personal information. In clinical trials the primary way in which researchers demonstrate respect is by asking for consent for participation.

Participant Information Sheets and Consent Forms

The National Statement requires that:

> *"Participation is voluntary and based on sufficient information requires an adequate understanding of the purpose, methods, demands, risks and potential benefits of the research"*

In the context of writing participant information sheets for clinical trials where participants' data will be linked the National Statement section 3.1.43 states:

> *"Where research involves linkage of data sets with the consent of participants, researchers should advise participants that use of data or information that could be used to identify them may be required to ensure that the linkage is accurate. They should also be given information about the security measures that will be adopted, for example the removal of identifiers once linkage is completed."*

More specifically it would be reasonable to include the following information:

- The fact that participants' data will be linked to data about them from other sources
- What other sources of data will be linked
- How and who will conduct the linkage
- What the benefits of linking their data will be
- Any risks associated with the linkage and analysis of the linked data and how you will minimise them

The language chosen to convey this information should take into consideration section 5.2.17 of the National Statement. It must be tailored to the needs of the potential participants i.e. the language and format should be appropriate for the cultural background, educational background and level, age and visual, hearing or communication impairments of the potential participants.

The NSW Population & Health Services Research Ethics Committee provides an example of wording about data linkage in a PICF.

See www.cancer.nsw.gov.au/data-research/research-ethics/submissions.

Access to some data collections will have specific legal and/or policy requirements around consent. For example, MBS and PBS data. Consent for MBS and PBS data requires a specific and separate consent form. Please refer to the Department of Human Services for further guidance. www.humanservices.gov.au/organisations/about-us/statistical-information-and-data

To obtain the current version of the MBS/PBS consent form, please email:

statistics@humanservices.gov.au

Access without Consent

In most cases it will not be justified to waive the requirement for consent to access linked data in clinical trials as participants can be asked when they consent to participation in the rest of the trial. In some circumstances it can be justified to use linked data in research without consent. For example if it is essential to include the whole population in a study or the results will not be valid. The

National Statement recognises this and includes two approaches (the opt-out approach and waiver of consent) for circumstances where gaining informed consent is neither practical nor feasible. Both these approaches raise important ethical considerations and also have legal significance.

Key Ethical Issues

Justice

The National Statement requires HRECs to take into account whether the benefits and burdens of research are fairly distributed and whether participants are being treated fairly. In some study designs e.g. registry-base randomised trials, the use of linked data lends itself to addressing the concerns associated with justice in research by enabling the inclusion of data from the whole population. Standard clinical trial designs can exclude the most disadvantaged members of the community from research. Language difficulties, transitory accommodation, mental illness and other health issues may all be obstacles for inclusion in a clinical trial. Study designs using population-based linked data are not only more inclusive, representative and unbiased but are also more just in their distribution of the benefits and burdens of the research than conventional designs.

Respect

The value of respect includes respecting people's privacy and their autonomy. This value, as described above, is usually demonstrated by only using people's information in research if they have made an active and voluntary choice to consent to participate.

In circumstances where using an opt-out approach or a waiver of consent researchers are required to:

- Justify the need to use one of these approaches by addressing the criteria in the National Statement (NS 2.3.5; 2.3.6; 2.3.9; 2.3.10)
- Demonstrate how they will show respect for the participants in the absence of informed consent.

Respect for participants can be demonstrated in a number of ways including:

- Involving the community who may be affected by the research in all stages of the development and implementation of a project and the dissemination of the results. For more information on community involvement in research see www.involvingpeopleinresearch.org.au/
- Ensuring that the results of the research are translated into improved practice and services
- Ensuring that information about the research and its results are made publicly available in language and in forums that make them accessible to the general community.

The Legal Framework

In addition to meeting the ethical standards outlined in the National Statement the use of linked data without consent must comply with legal requirements. The use of data without consent for research in law generally includes using an opt-out approach.

It is important that researchers understand the legal framework applicable to data linkage to ensure that all legal responsibilities are met.

The data linkage process is regulated by three bodies of law:

- Legislation empowering and regulating the collection, use or disclosure of information by government entities, such as public health statutes
- The common law of duty of confidentiality, and
- Privacy legislation.

In addition, researchers must comply with any contractual obligations relating to privacy and confidentiality e.g. confidentiality agreements. (4)

The bodies of law that relate to your research depend on the nature of the information, whether it is identifiable at any stage of the research project and the institution, jurisdiction and data collections involved.

The more data collections and jurisdictions involved in your project the more legislation you will need to comply with. The legal requirements for each project will also overlap with the ethical requirements. (5)

In order to ensure that you are aware of all your legal responsibilities it can be helpful to map the relevant legislation against the steps in the data flow diagram that you developed for your data management plan. You will need to assess the identifiability of the data at every step in your project to determine whether the data is 'personal information' and therefore which legislation applies.

Table 1: Mapping data flows against the legislation

Step	What kind of organisation? (private, state govt, Commonwealth govt)	Collection, use or disclosure of data?	Personal information or not?	Health information or not?	What legislation applies?

For further information about your legal responsibilities for your clinical trial you may need to seek legal advice.

Specialist Human Research Ethics Committees

Data linkage projects are currently excluded from the National Mutual Acceptance Scheme. In some jurisdictions research projects requesting access to government administrative data collections or

using that jurisdiction's data linkage unit are required to get ethics approval from a specialist HREC. This is usually the health department HREC. The client services officers at the relevant data linkage unit will be able to give specific advice about which local HREC your application should be sent to. Below is a list of HRECs which specialise in research using linked data.

Table 2: Australian HRECs specialising in data linkage review

Name of HREC	Contact
ACT Health Human Research Ethics Committee	ethics@act.gov.au
Australian Institute of Health and Welfare Ethics Committee	ethicssec@aihw.gov.au
NSW Population and Health Service Research Ethics Committee	CINSW-Ethics@health.nsw.gov.au
SA Department for Health and Ageing Human Research Ethics Committee	HealthHumanResearchEthicsCommittee@sa.gov.au
Tasmania Health & Medical Human Research Ethics Committee	human.ethics@utas.edu.au
Department of Health WA Human Research Ethics Committee	hrec@health.wa.gov.au

In Queensland, researchers can use any QLD Health or university HREC registered with the NHMRC.

The Centre for Victorian Data Linkage will accept a review from any recognised NHMRC approved HREC. There is an additional form specifically for Victoria (Victorian Specific Module) that is required in addition to the Online Form.

Note that clearance from the researcher's home institutional ethics committee is required before the AIHW Ethics Committee will assess the application.

If you are conducting a clinical trial in an Aboriginal or Torres Strait Islander community or you are planning to collect and/analyse your cohort on the basis of Aboriginal or Torres Strait Islander status you will need to adhere to the *"Ethical conduct in research with Aboriginal and Torres Strait Islander Peoples and communities: Guidelines for researchers and stakeholders"* 2018. Advice about how to put the principles in the Guidelines into practice can be found in "*Keeping research on track II*" 2018.

Some jurisdictions require review by a specialist Aboriginal and Torres Strait Islander HREC in addition to review by one of the specialist data linkage HRECs listed above.

Table 3: Aboriginal and Torres Strait Islander HRECs

Name of HREC	Contact

AH&MRC HREC (NSW)	ahmrc@ahmrc.org.au
Aboriginal Health Ethics Committee (SA)	Gokhan.Ayturk@ahcsa.org.au
Central Australian Human Research Ethics Committee	cahrec@flinders.edu.au
Human Research Ethics Committee of the Northern Territory Department of Health and Menzies School of Health Research	ethics@menzies.edu.au
WA Aboriginal Health Ethics Committee	ethics@ahcwa.org

References

1. National Statement on Ethical Conduct in Human Research, (2007 (Updated May 2018)).
2. Service AND. ANDS Guide: Data Management Plans. Australia: Australian National Data Service; 2017.
3. Michener WK. Ten Simple Rules for Creating a Good Data Management Plan. PLoS Comput Biol. 2015;11(10):e1004525.
4. Legal Responsibilities. Australia: Population Health Research Network; 2019 [Available from: www.phrn.org.au/for-researchers/roles-and-responsibilities/legal-responsibilities/.
5. Flack FS, Adams C, Allen J. Authorising the release of data without consent for health research: The role of data custodians and HRECs in Australia. The Journal of Law and Medicine. 2019;26:655-80.

CHAPTER 4: USING LINKED MBS AND PBS DATA IN CLINICAL TRIALS

Dr Anna Kemp-Casey

Objectives

By the end of this chapter you should be able to:

- Give examples of how linked MBS and PBS data can be used in clinical trials
- Describe the application process for access to linked MBS and PBS data with and without consent.

Introduction

The Medicare Benefits Schedule (MBS) and Pharmaceutical Benefits Scheme (PBS) datasets contain detailed and reliable information about an individuals' health care and can be a rich supplement to clinical trial data. When using any pre-existing data, it is important to understand the rules of the system in which they were collected and the purpose for which they were collected. This particularly applies to MBS and PBS datasets, which are complex and can contain different population and variable subsets.

Purpose and description of datasets

MBS

The MBS was established in 1984 to provide all Australians with access to health care. The MBS dataset captures information about a list of subsidised medical services. These services may be provided by medical doctors and allied health professionals in:

- the community (outpatients), and
- hospital to private inpatients.

The MBS dataset does not capture services provided to:

- public inpatients
- services not listed on the MBS (e.g. remedial massage)
- services which are on the MBS but where the patient does not meet the criteria for subsidy (e.g. rhinoplasty is only subsidised for patients with a significant acquired, congenital or developmental deformity).

PBS

The PBS was established in 1948 to provide Australian residents with access to affordable prescription medicines. The PBS dataset captures information about a list of prescription medicines dispensed from:

- community pharmacies,
- hospital pharmacies to private patients (inpatient or outpatient), and
- hospital pharmacies to public patients (upon discharge or to outpatients), except in New South Wales and the Australian Capital Territory.

The PBS dataset does not capture dispensing of:

- over-the-counter medicines (e.g. paracetamol, cough and cold preparations)
- prescription medicines not listed on the PBS (includes many new and novel therapies, such as alemtuzumab IV infusion for multiple sclerosis)
- medicines which are on the PBS but where the patient does not meet the criteria for PBS-subsidy (e.g. alendronate is only subsidised for the treatment – not prevention – of diagnosed osteoporosis), and
- medicines supplied to public inpatients.

Both datasets

The MBS and PBS are administrative datasets which primarily function as a record of payments from the Commonwealth to hospitals, pharmacists and patients. These datasets were not designed for research purposes and therefore do not capture key information researchers would like such as diagnosis/indication, intended doses for medicines, or test results.

Who is covered in the datasets?

The MBS and PBS datasets cover anyone with Medicare card: generally, all Australian citizens and permanent residents. The only exceptions to this are people serving in the armed forces and those in prison as the health care of these individuals are paid for by the Department of Defence and State Departments of Justice, respectively.

One of the main advantages of the MBS and PBS datasets for researchers is that they provide whole-population coverage. They are among the few schemes in the world to do this. Many other international health care schemes are only available to specific groups of people (e.g. low-income earners) and the datasets therefore cover non-representative sections of the population.

How to find and use MBS codes

MBS codes are numeric and range in length from one to five digits. There is no significance associated with the number assigned to a particular service; for example, five-digit codes are not necessarily more expensive or provided by a different specialty than those with two-digit codes. The current MBS schedule is available at http://www.mbsonline.gov.au/internet/mbsonline/publishing.nsfContent/Home (PDF for download) or searchable by key word or item code online at http://www9.health.gov.au/mbs/search.cfm.

The MBS is updated quarterly, with historic schedules available at www.mbsonline.gov.au/internet/mbsonline/publishing.nsf/Content/downloads. Items are added (listed) and removed (delisted) with each update. It is critical that you check schedules at the beginning, middle and end of your study period to ensure that you haven't missed any codes that were added or removed over this time. One way to do this is to use one of the category systems within the MBS such as broad type of service (e.g. non-referred attendances), or subgroup (e.g. Ophthalmology). This can help you ensure that you do not overlook a historic MBS code.

How to find and use PBS codes

PBS codes are alpha-numeric and are six characters long (five numbers and one letter). Newer items start with 1 and older items start with 0; however, the trailing 0 may not be present in all data cuts (e.g. 2013Y is the same as 02013Y). Each strength and preparation of a medicine has a different item code, but item codes do not differentiate between brands of medicines. The current PBS schedule is available at www.pbs.gov.au/browse/downloads (PDF for download) or searchable by key words or item code at www.pbs.gov.au/pbs/home.

PBS schedules are updated monthly, with historic schedules available at www.mbsonline.gov.au/internet/mbsonline/publishing.nsf/Content/downloads. Items are listed and delisted with each update, and item restrictions may also be changed. It is critical that you check the schedule at the beginning, middle and end of your study period to ensure that you haven't missed any codes that were added or removed over this time.

ATC codes were developed by the World Health Organization (WHO) and are used internationally to classify medicines by their generic name[1]. ATC codes are hierarchical and reflect: the anatomical system affected (A), therapeutic target (T), and chemical group (C) of the medicine. Unlike PBS item codes, ATC codes do not differentiate between strengths and preparations of a generic medicine. The main advantage of using ATC codes in analysing PBS data is that they are more stable (only occasional reclassifications occur)[1]; allowing researchers to select all PBS items which fall within a particular ATC code (or range of codes). This greatly reduces the likelihood of overlooking a historical PBS code. A PBS-ATC mapping table is publicly available[2].

How can the datasets supplement clinical trial data?

MBS and PBS data can supplement clinical trial data in two key ways:

1. Pre-trial patient history
   It is often desirable to collect detailed information about the care participants have received in the years prior to trial recruitment. MBS and PBS data provide major advantages over self-report or clinical chart review. Participant recall can be poor for the timing (or fact) of comorbid diagnoses, names of medications, duration of therapy and so on. These data may assist in decisions about trial eligibility and determining whether randomisation was successful. Chart

review is more reliable than participant recall but requires considerable time and resources from researchers and assumes all relevant medical records will be available.

2. Follow up
   Follow up may extend for weeks or years after an intervention. The use of MBS and PBS data to collect follow up information can reduce participant burden and researcher resources and can be used to assess factors such as medication adherence and clinical outcomes.

The MBS and PBS datasets do not directly capture information regarding diagnoses/indication or clinical outcomes. However, you can make reasonable inferences about these in some cases. Some health services and medicines make good proxies for some diagnoses or outcomes – either because they have no other known uses, or due to subsidy restrictions. For example, breast radiotherapy is an excellent proxy for breast cancer; as there are no other known uses for this treatment. By contrast, dispensing of an antidepressant is not a reliable proxy for a depression diagnosis as there are many approved and off-label uses for this class of medicines. The utility of any given proxy needs to be determined or justified on a case-by-case basis.

Strengths and limitations

Strengths

- Full population coverage – data highly likely to exist for all your clinical trial participants
- National coverage – data will be captured for your participants even if they have moved or undergone treatment in different jurisdictions
- Objective, highly reliable capture of services and medicines that were administered, and when they were administered

Limitations

- Not all services and medicines provided to patients are included in the data; just the subsidised ones
- No diagnosis information is captured about participants, but you can make assumptions about some items
- No clinical outcomes are captured, but you can make assumptions about some items
- MBS data capture varies by State/ Territory because jurisdictions vary in what they consider an inpatient procedure (inpatient data not collected for public patients)
- PBS data capture varies by State/Territory – know when your jurisdiction started contributing hospital data to the dataset

Variables available in the datasets

Most commonly used MBS variables

- Patient ID (deidentified)
- Patient sex
- Patient date of birth (may be partial)
- MBS item code
- Date of service
- Date of processing
- Hospital indicator flag
- Broad type of service
- Provider identifier number (deidentified)
- Provider specialty
- Provider geographic location (usually ARIA)[3]

Most commonly used PBS variables

- Patient ID (deidentified)
- Patient sex
- Patient date of birth (may be partial)
- Beneficiary type
- PBS item code
- Generic name
- Brand name
- ATC code
- Quantity supplied
- Repeat/original
- Date of service
- Date of prescription
- Date of processing
- Prescriber identifier number (deidentified)
- Prescriber specialty
- Prescriber geographic location (ARIA)[3]
- Pharmacy geographic location (ARIA)[3]

Data dictionaries: where to find them

The first data dictionaries were developed by the Commonwealth in 2015 to accompany a particular data extract provided to the States and Territories for the 2013-2014 financial year. To date these are the only data dictionaries which exist for the MBS and PBS datasets. These are therefore the best resource for determining which variables are available; however variable names and formats may differ in the extract you ultimately receive.

The MBS data dictionary is available at: meteor.aihw.gov.au/content/index.phtml/ itemId/603356.
The PBS data dictionary is available at: meteor.aihw.gov.au/content/index. phtml/ itemId/ 602524.

Beware of differences in names and formatting

MBS and PBS data are intentionally separated within the Department of Health. Data are stored in different buildings and maintained by different staff. In addition to this, there is considerable staff turnover within the Department. This means it is highly likely that any given data cut you receive will have been performed by a different person and contain different names and formats than previous extracts.

Important dates

The MBS dataset contains two key dates for each encounter:

i) the date the service was provided (*date of service*), and
ii) the date the service claim was processed by the Department of Health (*date of processing*).

The PBS dataset contains three key dates:

i) the date the prescription was written by the prescriber (*date of prescription*),
ii) the date the prescription was filled at the pharmacy (*date of supply*), and
iii) the date the prescription claim was processed by the Department of Health (*date of processing*).

It is critical to understand the difference between these dates, and which one/s was provided in your data cut. There are occasions where the above dates will be identical; however, they can also vary by months.

Important historical changes in the PBS dataset

Beneficiary status and 'below co-payment' data

There are two main categories of consumers covered by the PBS; concessional beneficiaries and general beneficiaries[4]. Eligibility for concessional status is largely determined by income and this group includes aged pensioners, Commonwealth Seniors Health Card holders, disability pensioners, single parent pensioners, and Health Care Card holders[4]. As a group, concessional beneficiaries tend to be older, sicker and poorer than the rest of the population[5]. Concessional beneficiaries constitute a small proportion of the population but account for the majority of medicine use in the

community[5]. Members of the community not qualifying for concessional status are general beneficiaries.

The amount a consumer pays for PBS medicines, referred to as a co-payment, is affected by an individuals' beneficiary status. Co-payments typically increase every year and the consumer is required to pay either the co-payment or the full cost of the medicine, whichever is cheaper. Concessional beneficiaries pay a lower co-payment than general beneficiaries; $6.50 and $40.30, respectively, as of 1st January 2019[4].

It is important to note that co-payments influenced PBS data capture until recently. Prior to April 2012, the PBS dataset only captured medicines for which a benefit was paid, i.e. the medicine price was higher than the co-payment. Consequently, medicines priced under the co-payment before April 2012 are missing from the dataset. This has the greatest impact on medicines dispensed to general beneficiaries because of their higher co-payment threshold. For some medicines, prices fluctuated to the extent that these items came in and out of data capture in the pre-2012 period. It is critical that researchers be mindful of the impact that below co-payment medicines will have on their data for medicines dispensed before April 2012. Previous PBS schedules should be examined to determine the price of medicines under study during the years of interest[6]. Researchers have published advice regarding analysis of below co-payment PBS medicines[7].

Capture of hospital dispensing to public patients

The PBS dataset captures includes medicines supplied to public patients at discharge from hospital. This is the case for all jurisdictions except New South Wales and the Australian Capital Territory. The other States and Territories began to contribute these records to the PBS dataset at different times[8]:

- Victoria: September 2001, amendment July 2003
- Queensland: August 2002
- Western Australia: 2002
- Northern Territory: January 2007
- South Australia: August 2008
- Tasmania: December 2010

If your study period precedes full PBS data capture in your jurisdiction, you will need to consider the impact when interpreting your data.

Capture of Section 100 medicines

Section 100 medicines are a group of highly specialised treatments which are typically administered through hospitals (e.g. chemotherapy). Historically, these medicines were bulk-processed by the Department of Health at the end of each month and were therefore not attributed to individual patients within the PBS dataset. This practice was modified to allow patient-level data capture of Section 100 medicines for private hospitals in December 2011 and public hospitals in April 2012. Consequently, PBS data preceding these dates do not contain dispensing of Section 100 medicines.

Applying for MBS/PBS data

There are two approved bodies which can facilitate linkage of clinical trial data with MBS and PBS data:

1. Population Health Research Network www.phrn.org.au/for-researchers/data-access/
2. Australian Institute of Health and Welfare www.aihw.gov.au/data-linking/

The websites for both organisations provide detailed advice about the application process.

Informed consent is a key principle of ethical research. The NHMRC Statement on Ethical Conduct in Human Research[9] states that consent is required for participation in medical research except under particular circumstances.

Obtaining participant consent is standard for clinical trials, and it is most practical to obtain consent to access MBS and PBS data at the same time as trial consent is sought. Be aware the Department of Health have very specific requirements for the wording of consent forms relating to MBS and PBS data, and it is unlikely that a modification of existing clinical trial information and consent forms will be adequate. Contact the Department of Health to obtain their most recent participant information and consent templates, as these are updated routinely.

There are circumstances where obtaining participant consent is problematic or impossible. For example, clinical trial participants may have died in the period between the intervention and when consent for MBS and PBS data is being sought. For instances such as this, it is necessary to obtain a waiver of consent. A waiver of consent to use personal medical information can only be granted by an approved Human Research Ethics Committee (HREC). In order to grant a waiver, the HREC must be satisfied that:

a) involvement in the research carries no more than low risk to participants (see paragraphs 2.1.6 and 2.1.7 of the National Statement on the Ethical Conduct of Human Research)
b) the benefits from the research justify any risks of harm associated with not seeking consent
c) it is impracticable to obtain consent (for example, due to the quantity, age or accessibility of records)
d) there is no known or likely reason for thinking that participants would not have consented if they had been asked

e) there is sufficient protection of participant privacy
f) there is an adequate plan to protect the confidentiality of data
g) in case the results have significance for the participants' welfare there is, where practicable, a plan for making information arising from the research available to them (for example, via a disease-specific website or regional news media)
h) the possibility of commercial exploitation of derivatives of the data or tissue will not deprive the participants of any financial benefits to which they would be entitled
i) the waiver is not prohibited by State, federal, or international law.

Analysing MBS and PBS Data

While the MBS and PBS datasets are a rich source of data, they are complex datasets to work with. The following are highly recommended while you are planning your study are before you commence analysis:

A clear research question

A clearly articulated research question is vital to making the right choices about the MBS and PBS codes you are including in your study, and the time periods of interest. The sheer scale of MBS and PBS data mean that 'fishing expeditions' are likely to lead to statistically significant, but clinically meaningless, associations.

Time

Be aware that the time between request and receipt of linked MBS and/or PBS data can be lengthy, (>1 year) and plan accordingly. The time-consuming component of analysis with these datasets is not the statistical component (e.g. the Cox regression) but the data preparation. It takes many months to take multiple datasets with one record per health encounter (e.g. potentially hundreds of records per person) and convert them into one dataset with all the relevant information on one record per person. This is where 95% of the work occurs.

Syntax writing

It doesn't matter which statistical package you use (many researchers use SPSS, R, SAS, and Stata), but it is important that you use syntax/code rather than drop down menus. The steps involved in creating an analysable dataset are long and complex, and often need to be refined and repeated dozens of times. If each step is being done manually this becomes a painful and lengthy process. Syntax writing is also useful as a detailed record of the decisions made regarding codes, time periods, study inclusions and exclusions, and the way variables were categorised. These details are difficult to remember years later when journal reviewers have questions.

Friends

Cultivate friendships with a wide range of professionals including clinicians, pharmacists, clinical coders, and consumers. It is unlikely that any one person will have the full range of expertise and lived experience needed to interpret these complex data alone.

References

1. WHO Collaborating Centre for Drug Statistics Methodology. ATC/DDD Index 2016. WHO, Oslo, Norway; 2016. http://www.whocc.no/atc_ddd_index/ (accessed 17 Jan 2016).
2. Department of Health. PBS and RPBS section 85 date of processing and date of supply data. Department of Health, Canberra; 2015. www.pbs.gov.au/info/statistics/dos-and-dop/dos-and-dop (accessed
3. Hugo Centre for Migration and Population Research. ARIA and Accessibility. University of Adelaide, Adelaide; 2017. http://www.adelaide.edu.au/apmrc/research/projects/category/aria.html (accessed June 29 2017).
4. Department of Health. Pharmaceutical benefits: fees, patient contributions and safety net thresholds. 2019. http://www.pbs.gov.au/info/healthpro/explanatory-notes/front/fee (accessed 29 March 2019).
5. McManus P, Birkett DJ, Mant A. Prescription use by patients with concession cards. *Med J Aust* 1998; **169**: 285-286.
6. Department of Health. PBS Publications. 2016. http://www.pbs.gov.au/browse/publications (accessed Jan 21 2016).
7. Paige E, Kemp-Casey A, Korda R, et al. Using Australian Pharmaceutical Benefits Scheme data for pharmacoepidemiological research: challenges and approaches. *Public Health Research & Practice* 2015; **25**: e2541546.
8. Department of Health. Appendix I: Public Hospital Pharmaceutical Reforms: PBS and hospitals. Department of Health Canberra; 2013. http://www.health.gov.au/internet/main/publishing.nsf/Content/chemotherapy-review/$File/appendix-i.pdf (accessed 29 March 2019).
9. National Health and Medical Research Council, Australian Research Council, Universities Australia. National Statement on Ethical Conduct in Human Research 2007. Canberra: Commonwealth of Australia, 2007 (updated 2018).