Queensland Data Linkage Framework

February 2017

Queensland Government

Queensland Data Linkage Framework

Published by the State of Queensland (Queensland Health), January 2016

For more information contact:
Statistical Analysis and Linkage Unit, Statistical Services Branch, Department of Health, GPO Box 48, Brisbane QLD 4001, email HSBresearch@health.qld.gov.au, phone (07)32340128.

An electronic version of this document is available at http://www.health.qld.gov.au/hsu/pdf/other/qlddatalinkframework.pdf

Contents

Introduction

Purpose of this document

This document details the principles, structure and processes of the Queensland Research Linkage Group (RLG) located within the Statistical Services Branch at the Queensland Department of Health, which is part of the Queensland node of the Population Health Research Network (PHRN).

Context

Data linkage is a process by which information within or across multiple sources that relates to an individual entity can be combined. There has been increasing interest in recent years in utilising data linkage techniques to link data within and across the extensive range of population-based and health service data sets that exist in Australia at a federal and state level. Data linkage is an efficient way to enhance existing data to increase its utility for informing population health and clinical research as well as policy development and performance measurement.

Since the 1970s Western Australia has developed substantial expertise and capability in data linkage. This culminated in 1995 with the establishment of the Western Australia Data Linkage System (http://www.datalinkage-wa.org/). The Western Australia Data Linkage Unit (WADLU) has over time built a core linkage capability from statutory elements including birth and death registrations, and hospital separations. There has been a growing momentum within the Australian context to learn from and leverage the Western Australian experience and capability.

Within this context, the Australian Government has provided funding for a number of National Collaborative Research Infrastructure Strategy (NCRIS) initiatives. One initiative has been to establish the Population Health Research Network (PHRN). This is a national network with representation from the Commonwealth and all States and Territories that aims to develop data linkage infrastructure and capability and progress linkage across Australia's extensive health data collections to facilitate population health research.

The Queensland node

In response to expressions of interest for the PHRN a consortium was formed in Queensland with representation from Queensland Health, the CSIRO Centre for eHealth Research, the University of Queensland, Griffith University, James Cook University and Queensland University of Technology. This group of stakeholder worked to request NCRIS funding and to set up data linkage services within the Queensland Department of Health and continue to provide input into the services provided through the Queensland Data Linkage Reference Group.

Data linkage services for the Queensland node are provided by the Research Linkage Group (RLG) at Queensland Health. Locating the linkage unit within Queensland Health enables linkage services to be conducted in a secure environment, ensuring compliance with strict security, privacy and confidentiality requirements. While in the future there may be a small charge for research requests, funding under the NCRIS/PHRN and from Queensland Health currently subsidises data linkage work

conducted within the RLG which includes systematised linkage within Queensland Health's major data collections and customised linkage and/or provision of linked data for internal and external analysis and research.

What is data linkage?

Data linkage allows for the identification of distinct entities within datasets and between datasets. For example, data linkage can be used to identify the number of times a person is admitted to hospital in a year or mortality following hospital discharge. Linkage uses identifying patient data such as names, date of birth and address, but these data are accessed solely for the purposes of linkage. In order to protect patients' privacy and confidentiality these data items are never released to researchers; instead, project-specific patient identifiers are created for release to researchers to allow them to determine which data relate to an individual and thereby to combine data from different sources.

Benefits of data linkage

There is a growing recognition of the value of data linkage for population based health research as well as for health service policy and planning. The benefits of a comprehensive data linkage system include:

- Improved cost effectiveness of health research – linking existing data (that is often routinely collected) is a relatively cheap and effective alternative to actually conducting large scale longitudinal research studies/or clinical trials or to collecting extra data to supplement an existing database when that data exists elsewhere.
- Data linkage enables the re-use of existing data sets and adds value through improved data quality and analysis.
- Improved collaboration and health research outcomes – linking data from multiple data sources requires collaboration and negotiation across multiple stakeholders. This input and the resulting availability of linked data further promotes and adds to increased collaboration and research outputs.
- Improved management of communications – for example linking patient cohort data to death data can avoid sending requests for information to (the families of) deceased persons.

Analyses of linked data have informed population health and health services evaluation and research - in areas as diverse as statistical process control for clinical improvement, chronic disease management, the interface of health and emergency services, use of hospital and community-based mental health services and patient satisfaction.

Methods of linking

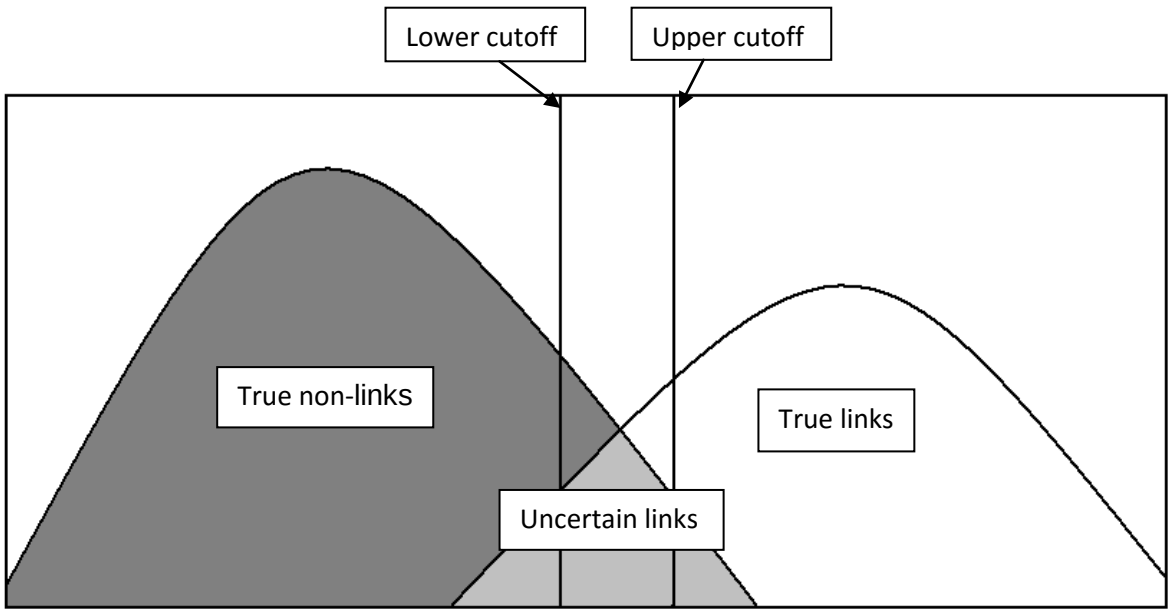
In Queensland, both deterministic and probabilistic methods of linking records are used. Both methods have their own advantages and their usage is dependent on the information available.

Deterministic linkage involves the linking of data sets using unique identifiers such as a patient/client unique identifier or through comparing fields such name, street name, year of birth, street number with the requirement that the records agree on all characters. Deterministic linking (sometimes called exact matching) is not without error. However, with the use of computer programs and partial identifiers such as postcodes the proportion of true matches can be increased.

Probabilistic linkage involves the use of statistical models and mathematical formulae (algorithms) to estimate the probability of data from different data sets having commonality (e.g. the same person/event). Matching variables are assigned weighted scores so, for example, rare surnames are given a higher weight than common surnames. Additionally names are converted to a phonetic code (soundex/NYSIIS) in order to handle spelling discrepancies (e.g. Mcdonald vs. Macdonald, Smyth vs. Smith). Dates of birth that do not match exactly are still given some weight if there is a viable rearrangement or substitution of dates. The main advantage of this method is that data from different sources, and of varying quality, are able to be linked successfully whereas deterministic linkage may fail to identify many true matches due to minor differences.

Clerical review is used to manually inspect the 'grey area' of uncertain matches in probabilistic linkage. When pairs are ranked by total weights from linkage, there will be a lower cut off below which pairs are considered non-matches, and an upper cut off above which pairs are considered true matches. Between these bounds lie the paired records that are a mixture of true and false matches, and human judgement is required to decide whether to link them.

A typical distribution of true and false links is demonstrated below – the overlap can be described as the 'grey area':



Data linkage and patient privacy

Privacy and confidentiality of the Statistical Services Branch's data holdings are covered by Part 7 of the Queensland Hospital and Health Boards Act 2011 and relevant sections of the Public Health Act 2005 and the Private Health Facilities Act 1999, and the RLG has an agreed approval process for linkage projects. Both the linkage and analysis environments are protected by firewalls and access audit trails are maintained.

Requests for linked data must be made in accordance with Queensland Health's data release protocols which are covered in detail under 'Protocol for researchers requesting data'. The procedure to access data depends whether the requestor is employed by Queensland Health or is external, and the purpose of the request e.g. research, health service improvement.

Data linkage in Queensland

Datasets for linkage

The most common datasets used for linkage include:

- Queensland Hospital Admitted Patient Data Collection (QHAPDC)
- Queensland Perinatal Data Collection (QPDC)
- Queensland Cancer Registry (QCR)
- Registrar General (RG) deaths*
- Emergency Department Information System (EDIS)
- Community Integrated Mental Health Application (CIMHA)

*For data linkage purposes, Queensland Health acts as proxy Data Custodian under a Memorandum of Understanding (MOU) with the Queensland Registrar General.

Further details about these datasets are available in Appendix 1.

There is also a Master Linkage File (MLF) containing permanently linked references to QHAPDC, QPDC, RG births, RG deaths, and EDIS. The use of this file saves a significant amount of RLG resources and results in a faster processing time, benefiting researchers. The MLF is now updated in near-real time such that data are extracted from all sources twice each month and linked with all other records included in the MLF. The currency of the data included in the MLF is limited by the currency of data that are available in the source data collections. See Appendix 1 for details of approximate lag times for submission of data for each data collection. Work is currently being undertaken to expand the MLF with data from further sources including the Notifiable Conditions System (NOCS), the Vaccination Information and Vaccination Administration System (VIVAS) and the Queensland Ambulance Service (QAS) to improve its coverage and therefore usefulness.

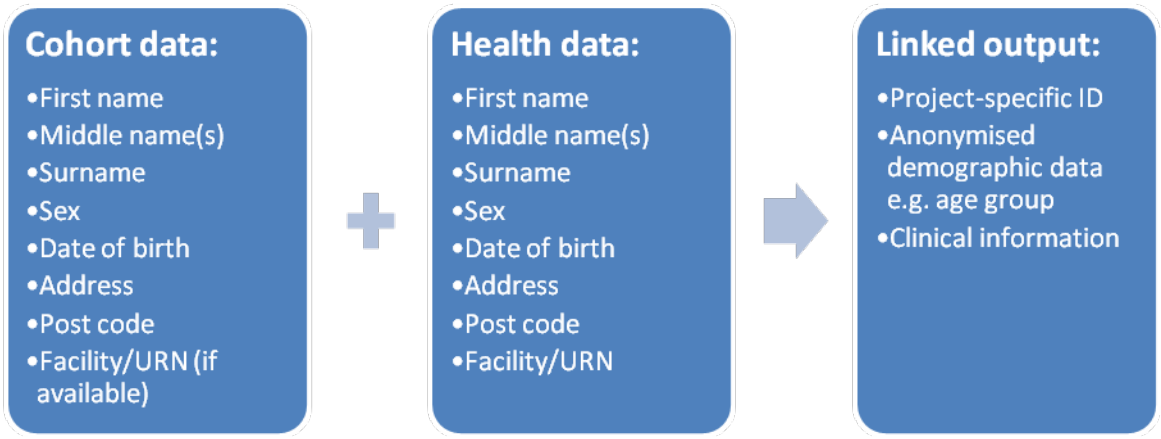
Current Master Linkage File coverage:

QHAPDC Hospital Admissions*	01 January 2004 to latest available**
Queensland Perinatal Data Collection*	01 July 2007 to latest available**
Registrar General births	01 July 2007 to latest available**
Registrar General deaths	01 January 2004 to latest available**
Emergency Department Information System	01 August 2011 to latest available**
Elective Surgery Waiting List	01 July 2014 to latest available
Outpatient Waiting List	01 July 2015 to latest available

**names and addresses for private facilities and in perinatal data were not supplied to the central data collection prior to 01 July 2007 so the quality of the linkage for these patients is questionable prior to this date and may result in bias in analysis undertaken using data from the period prior to 1 July 2007. For example, readmission rates for private hospitals may appear lower than for public hospitals due to failure to link data for private patients. For this reason, the MLF does not include data prior to 1 July 2007.*
***updated 15/02/17.*

Overview of the data linkage process

The figure below is a simplified overview of the linkage process involving linking a client's cohort file to Queensland Health datasets such as QHAPDC, resulting in an anonymised file containing only summarised clinical information relevant to the research proposal:



1. Client identifies the population of interest. This could be a supplied cohort group from the researcher or Queensland Health clinical database, or a specific group from a health-related data collection(s) e.g. admitted or emergency department heart failure patients transported by the Qld Ambulance Service.

2. Cohort file containing only identifiers for linkage plus a person and (if applicable) record ID sent securely to RLG. Linkage staff import, clean and standardise identifying demographic information for linkage. Issues such as missing data and duplicates are noted for reporting back to the client.

3. Linkage staff use most efficient/appropriate method of linking cohort to required dataset(s), for example if MLF can be used this will greatly reduce clerical checking of uncertain 'grey area' links. Similarly, for linkages involving QHAPDC if compatible identifying reference numbers are available in the cohort data, a deterministic matching exercise can be done, with any remaining unmatched cohort linked probabilistically. A project-specific linkage key is created during this process.

4. The project-specific linkage key is attached to the relevant clinical service records and required variables are extracted. Where necessary, data items are manipulated to meet agreed privacy protocols e.g. date of admission formatted as year/month, age transformed to 5-year bands.

5. The whole linkage process is quality-checked by a separate member of linkage team. Quality assurance is discussed in further detail below.

6. Methodology and any data issues are summarised in a report, along with output variable descriptions and details of what is contained in output worksheet(s)/file(s). Where possible output is provided in Excel, but large files may be supplied in SAS or as text files.

7. Final output and report are sent to the client via a secure method of transmission. Normally WinZip is used to password protect the file(s) with 256 bit encryption. If the client does not have use of WinZip an alternative encryption method is agreed. Passwords are always given separately by telephone.

Quality Assurance (QA) of the linkage

A number of checks are carried out at each stage of the linkage process, from receipt of cohort data (if applicable) to ensuring the final output meets the client's requirements and complies with privacy protocols.

A list of the Quality Assurance checks applied to linked data in Queensland is shown in Appendix 2. These checks are applied as an additional layer of checking following matching by linkage software and manual review of all grey area identified. These checks involve automated identification of any records or groups of records meeting the criteria and manual review to determine whether they should or should not be considered a match. The checks listed are used on the Master Linkage File and, where applicable, for linkage conducted for research requests. These checks are based on the list of quality assurance checks used by the Centre for Health Research Linkage in NSW (CHeReL) (see http://www.cherel.org.au/media/24160/qa_report_2012.pdf.), personal communication with other linkage units within Australia, and quality issues identified with Queensland data or with the linkage process and tools being used in Queensland.

Some examples of checks applied at different stages of the linkage process for a research request are described below:

Cohort data

- Check cohort dates of birth and (if applicable) clinic/procedure/other dates for follow-up are within range specified in the PHA application.
- Identify and quantify missing values in identifying variables.
- Check for duplicate records if cohort file should consist of distinct persons.
- Clean and standardise identifying variables to optimise consistent matching with health data.

During the linkage – 'grey area' checking

Weights assigned in the linkage process can to some extent automatically determine true positive and true negative links, but in the middle is an area of uncertainty that requires human judgement. Within this area, paired sets of identifying information are examined manually to determine whether or not they are true matches. Checks include:

- Are first names different when everything else matches? If so, is there a middle name in one record that matches first name in the other? If not, these may be twins so not linked.
- If dates of birth do not match, does it appear it is estimated in one dataset, e.g. 1 January of same year/one year out?
- Consider reasons surnames may not match. Differences in surname may be due to marriage or misspelling. Although less common than females, there are instances where males have changed surname, particularly children.
- Address changes are common, but geographical distance should be considered. Even when all other details match, it is less likely this is a true match if the name is common.

Also, the proportion of linked to non-linked cohort records will be considered. For example where 100% linkage is expected but not attained, unlinked cohort records will

be examined for characteristics such as missing/poor quality information and may be manually searched in the relevant health databases to establish a link.

Final output

- Check instances where date of death precedes other dates e.g. admission.
- Ensure that only one distinct cohort person has linked to each health record.
- Query oddities such as a woman apparently giving birth twice within six months.
- Where age is calculated check for unrealistic values e.g. 130 years old.
- Check for multiple death records linking to one person.

Limitations

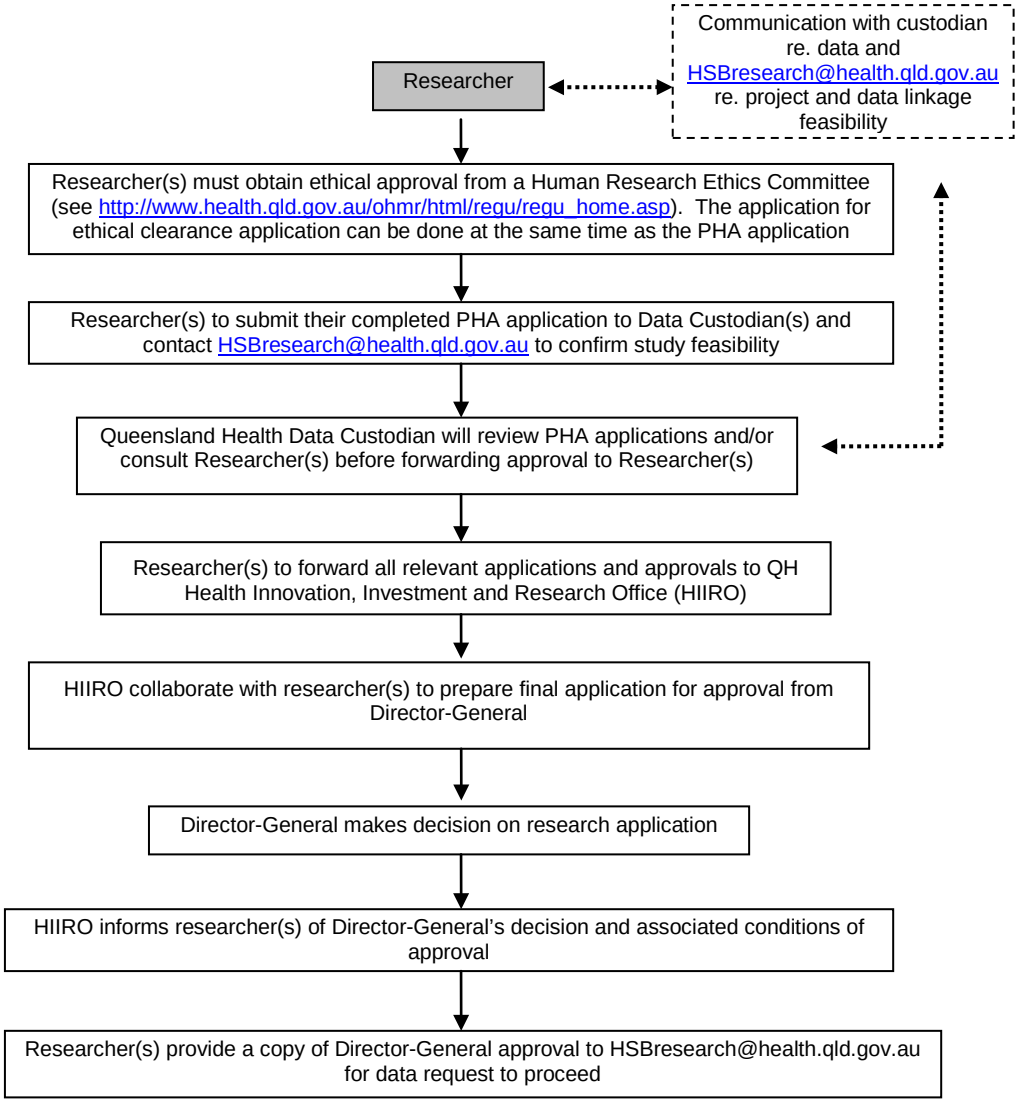
Although quality checks are carried out, it is important to bear in mind that there are limitations to data linkage within Queensland data sets that may make linkage output appear to be incorrect. For example, when hospitalisations are linked to death there may be instances when a discharge type of 'death' is not linked to an RG death record. This could be due to the death being registered in a different state – for example, residents of northern NSW may be admitted to facilities in southern QLD and vice versa.

QA for data linkage clients

Analysts working with linked output should also apply their own quality checks. CHeReL has produced useful guidance at this link: http://www.cherel.org.au/media/24762/edit_checks_-_Oct12.pdf. The guidance is equally applicable to Queensland linked data.

Protocol for researchers requesting data

Overview of Application Process for Access to Confidential Health Information held by Queensland Health for Research Purposes



Health Innovation, Investment and Research Office

The Health Innovation, Investment and Research Office (HIIRO, previously known as the Health and Medical Research (HMR) Unit) is the central port of contact for researchers seeking advice and direction on ethical and governance issues associated with the conduct of research using Queensland Health data.

Website http://www.health.qld.gov.au/ohmr/default.asp

Email pha@health.qld.gov.au

Any researcher applying for access to identifiable data held by Queensland Health for research purposes must obtain ethics approval for the research proposal as well as an approved application for data release under the Public Health Act (PHA) 2005.

Ethics and Governance

This involves:

a) Seeking ethical and scientific approval of the research protocol by a recognised Human Research Ethics Committee (HREC) using a Low/Negligible Risk Ethics Form or National Ethics Application Form (NEAF)

b) Completing the research governance component of a Site Specific Assessment (SSA) to determine the level of support and suitability of a research study to be conducted and completed at a site, whether that study is multi-centre or single-site.

Please refer to http://www.health.qld.gov.au/ohmr/html/regu/for_researcher.asp for more information.

Public Health Act application

To access identifiable or potentially identifiable information held by Queensland Health where researchers are unable to obtain participant consent, researchers need to complete a Public Health Act (PHA) application. This form covers the purpose of the research, the methodology of the study, the data required from Queensland Health data collections, how the privacy and confidentiality of the data released will be maintained securely, as well as data custodian sign-off. The Director-General (DG) or their delegate may grant access to health information for the purpose of research if it is in the public interest.

The PHA form requires the signature of each data custodian in order to be further processed by HIIRO. Once the custodian has signed the PHA form in section 10, the PHA form will be scanned and emailed in PDF form back to the researcher who then needs to forward all paperwork to the HIIRO.

The PHA is a legal document that authorises the release of the data items and time periods as specified and signed off. Following approval, if any amendments are made to the PHA application eg. request for additional data items, the form must be re-submitted to the relevant data custodian(s), and go through the entire approval process again.

Please refer to http://www.health.qld.gov.au/ohmr/html/regu/aces_conf_hth_info.asp for further information. Queries about the PHA application process can be emailed to HSBresearch@health.qld.gov.au.

Data custodian

It is strongly recommended that researchers contact data custodians directly as early as possible, even before ethics application, to determine if the data they require are available. **All researchers should also contact HSBresearch@health.qld.gov.au to ensure that the data linkage project they are proposing is feasible and that there are resources available to perform the data linkage.**

The Statistical Services Branch is the current data custodian for QHAPDC, QPDC and RG birth and death data, as well as the provider of linkage services. The Director of the Statistical Reporting Team is responsible for confirming that the requested data are available and able to be provided for a specific research project – this confirmation is given via section 10 of the PHA application.

The data custodian has the additional responsibility of ensuring that appropriate patient privacy is maintained when considering requests for data. For this reason, only those Queensland Health data that are clearly shown to be essential for the research project are considered for release.

Contact details for data custodians for Queensland Health data collections are available at https://www.health.qld.gov.au/__data/assets/pdf_file/0019/157123/data_custodian_list.pdf A list of data collections frequently requested for linkage with custodian details and links to information about those data collections is available at https://www.health.qld.gov.au/__data/assets/pdf_file/0032/644846/data_collection_table.pdf.

Application to Director General

HIIRO will prepare a brief of the research proposal and forward it to the Director-General (DG) or his Delegate for final approval. Once this final approval is granted, HIIRO will record the approval in the Research Registry and send a letter to the researcher advising of formal delegate approval. The researcher then needs to provide a copy of this notice from HIIRO to HSBresearch@health.qld.gov.au, who will then notify the RLG that the DG or delegate has granted approval for the data release.

Signed Consent

If a researcher is able to obtain signed participant consent from participants to use their personal or identifying information for a clearly specified research study, then the PHA application does not apply to health information held by QH, as the disclosure of information is authorised by another act (for example, Section 144 Hospital and Health Boards Act 2011). A copy of a blank consent form and information sheet, ethics approval and list of requested data items should be supplied to the data custodian. The data custodian needs to be satisfied that the consent form is relevant to the request and provides enough information for the participants to make an informed decision about consent. The custodian also has discretion to request a copy of signed consent forms. If your request involves linkage of data sets please email HSBresearch@health.qld.gov.au.

Penalties for misuse of linked data

The Australian Code for the Responsible Conduct of Research (2007) describes the process for handling breaches of the code including:

- Fabrication of results
- Plagiarism
- Conducting research without ethics approval

The Code is equally applicable to research involving the use of linked data as to any other method of research. Linked data must only be used for the research purposes outlined in the ethics and PHA applications as approved.

The full text of the Code can be found at the following link:
http://www.nhmrc.gov.au/_files_nhmrc/publications/attachments/r39.pdf

In addition, from 28 September 2016 changes to the Privacy Act, 1998 make it illegal to re-identify government data which has been “de-identified

Advice for application for linked data

In general, the release of data is guided by the principles outlined in the Guidelines for the Use and Disclosure of Health Data for Statistical Purposes developed by the National Statistical Information Management Committee in 2007. This document is available at http://www.health.qld.gov.au/performance/docs/simc_anonymisation.pdf and should be referred to when trying to determine what level of information to request.

Some points to consider when negotiating with data custodians or completing the PHA form:

- The scope of the data (eg public and/or private) and the variables required needs to be consistent with and justified by the purpose of the research. Requestors must specify the inclusion and exclusion criteria for the study group of interest and data range. For example, if only data for acute episodes for children under 10 years of age who reside in a specific area are required, then these criteria must be clearly stated in the PHA. Broad data requests (eg asking for all episodes of care for all ages) would not be considered an appropriate or targeted request for QHAPDC data for a specific research question;

- Additional justification for any fields may be requested where it is not clear why a field is required from the research objective/methodology described. For example, justification may be sought if all the additional diagnosis codes were requested if the research project is focussed on a specific health condition of interest;

- Check whether the data items or named information are available for the entire period requested eg QPDC mother names and QHAPDC private facility patient names were not collected until July 2007 (see Appendix 1 for further details regarding data availability);

- Ensure the conditions and/or procedures of interest are clearly defined. The relevant ICD diagnosis or procedure codes for the period of interest should be provided.

- RLG data released may be confidentialised to maintain patient privacy but the degree of detail in the data will be dependent on the objective of the research. For example:
    - age groups may be released instead of single year;
    - instead of listing each individual co-morbidity, a 'flag' could be used to denote that one of a group of specific diagnoses were recorded as a co-morbid condition;
    - procedure date may be reported as month/year instead of day/month/year;
    - geographical classification such as ARIA+ could be mapped instead of supplying postcode/SLA;
    - if names of both public and private hospitals are requested, the private hospitals will remain de-identified;
    - if intervals between events are of interest, these may be provided rather than the actual dates of events;

- It should be recognised that the term “de-identified” is used frequently in documents to refer to sets of data from which only names have been removed. Such data may remain “potentially identifiable”. See definitions in glossary, below.

- Consider how will the confidential information be disclosed between the data custodians and researchers? A method such as encryption and password protection with WinZip, with secure transfer of the WinZip file via email with the password communicated by phone on receipt of data is preferred.

- Review how the data will be stored. For example, a lockable safe with limited access to the office or secure university network server may be viable options. The data custodian should be assured that confidential patient information will be held securely and accessible by only those researchers named on the PHA application.

- For each input dataset to be linked, fields required for analysis should be listed separately from those identifying fields (full name, full address, date of birth, sex) required for linkage only (this includes any cohort datasets to be supplied by the researcher);

- What is the explicit scope of data required in the output dataset? When linking data from separate datasets, there will be ‘subsets’ of records created with records that appear in both data sets and records that belong to only one data set. For example:

Figure 1

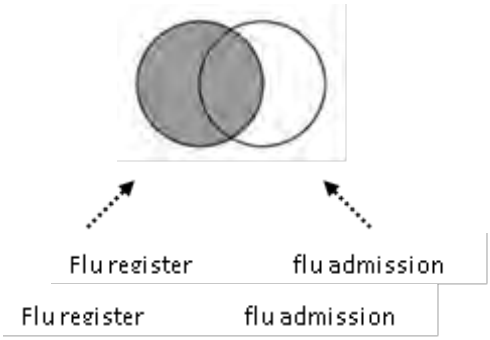
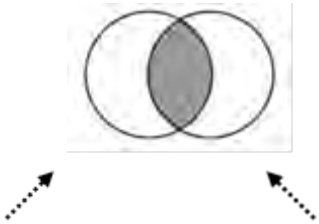


Figure 2



A hypothetical dataset contains patients on a flu register who received a vaccination, while another dataset may contain patients admitted to hospital for influenza. Figure 1 shows all patients on the influenza vaccination register plus any influenza-related hospitalisations, while Figure 2 shows only patients who were both on the influenza register and had an influenza-related hospitalisation, i.e. the patient appears in both datasets.

Is one report output table required or separate tables for each data source (e.g. ED presentation, hospital separation details, clinical)? Note that this may be dictated by the level of information involved.

Appendices

Appendix 1 – Metadata for commonly requested Queensland Department of Health data collections

Useful resources

Queensland Hospital Admitted Patient Data Collection manual:
https://www.health.qld.gov.au/hsu/collections/qhapdc

Queensland Perinatal Data Collection manual:
https://www.health.qld.gov.au/hsu/collections/pdc

Data Quality Statements: http://www.health.qld.gov.au/hsu/

Technical reports regarding some of the data quality issues to be aware of when analysing and interpreting Queensland administrative data collections are available at http://www.health.qld.gov.au/hsu/tech_report/tech_report.asp

The Statistical Services Branch is currently working towards publication of a full data dictionary on its website which will include further details regarding values for each data item, changes to data items over time and when data items were introduced. Some information is currently available at http://www.health.qld.gov.au/hsu/qhdd/QHDD-OurPerf.pdf or http://oascrasprod.co.health.qld.gov.au:7900/pls/qhik_prd/qhik_main_menu.entire_page and http://oascrasprod.co.health.qld.gov.au:7900/pls/crd_prd/f?p=144:1:2072367219751511::NO (currently available to Queensland Health staff only) and https://www.health.qld.gov.au/__data/assets/pdf_file/0032/644846/data_collection_table.pdhttps://www.health.qld.gov.au/__data/assets/pdf_file/0032/644846/data_collection_table.pdf

Scope and timeframes for commonly requested Queensland Department of Health data collections

Data Collection name	MLF coverage*	Date names and addresses available from and to**	Comments
Queensland Hospital Admitted Patient Data Collection	1 Jan 2004 (public); 1 July 2007 private. Includes latest available data.	1 July 1995 (public); 1 July 2007 (private). Available approx. 35 days after the end of the month when the patient was separated from an episode of care; validated data are available up to approx. 3 months prior to current date. Timeliness and completeness of data may vary by facility.	ICD-10-AM was introduced 1 July 1999 in Queensland Hospitals. Data requests for diagnosis information prior to this need to include ICD-9-CM codes.
Queensland Perinatal Data Collection	1 July 2007 to latest available data.	1 July 2007. Available up to approx. 9 months prior to current date, though timeliness and completeness of data may vary by facility.	It is possible to obtain identifying details for mothers prior to this for public hospitals via admitted patient data.
Death Registration Data	1 Jan 2004 to latest available data.	1 July 1996. Data are updated weekly, though not all deaths are registered in the same week/month/year that they occur.	Data from 1980 is available electronically from the Registrar General's office, though there is generally a charge associated with such requests.
Coded Cause of Death Data (ABS)		NA – able to be merged onto death registration data via death registration id and year.	As at 5 Jan 2016, the most recent data coded with ICD-10 cause of death details is for the 2012 calendar year. Coded data are available from 1980.
Birth Registration Data	1 July 2007 to latest available data.	1 January 2007. Data are updated weekly, though not all births are registered in the same week/month/year that they occur.	Data from 1980 is available electronically from the Registrar General's office, though there is generally a charge associated with such requests.
Emergency Department Information System	1 August 2011 to latest available data. 1 July 2008-31 July 2011 will be added when time permits.	Data are available for all major hospitals from 1 July 2008. Data are collected in real time.	Data are available for selected hospitals prior to this, but not for all.

*Where data are already available in the MLF this will cut down on the time taken for a data linkage request.
**Note that not all data items in each collection will be available for the entire period where names and addresses are available. This information will be available in the on-line data dictionary when it becomes available. Currently this information is available at https://www.health.qld.gov.au/__data/assets/pdf_file/0032/644846/data_collection_table.pdf for selected data collections or from individual data custodians.

Appendix 2 – Quality checks performed by the Queensland Research Linkage Group

Error ID	Checks for false positive links	Automated check of MLF
FP01	Hospital admission or other event date > date of death (excluding organ procurement)	***Y***
FP02	Date of birth and first 4 letters of first name and first 4 letters of surname differ	***Y***
FP02A	Date of birth and first 4 letters of surname and postcode differ	***Y***
FP02B	Date of birth and first 4 letters of first name and postcode differ	***Y***
FP03	Hospital admission date is prior to birth date	***Y***
FP04A	Episode of care with separation mode=death and end date after registered date of death	***Y***
FP04B	Episode of care with separation mode=death and end date before registered date of death	***Y***
FP04C	Organ procurement date < date of death.	***Y***
FP05	Person ID contains multiple death registration records	***Y***
FP06A	Person ID contains multiple birth registration records	***Y***
FP06B	Person ID contains multiple admitted patient data birth records	***Y***
FP06C	Person ID contains multiple PDC birth records	***Y***
FP07	Person ID contains a QPDC Mother record or a QPDC Baby record or birth registration record and mother's age<13 years in year that birth was recorded.	***Y***
FP08	Person ID contains a QPDC baby record or mother QPDC or birth registration record and mother's age >50 years in year that birth was recorded.	***Y***
FP09	Person ID contains QPDC baby record and birth registration record and Mother's name does not match.	***Y***
FP10	Person ID has Death Date before QPDC mother baby's date of birth.	***Y***
FP12	Average number of records per month in MLF ID is 50% greater than 75th percentile (unexpectedly high number of records)	***Y***
FP12A	Average number of records per month in MLF ID is 50% greater than 75th percentile – chemotherapy patients only	***Y***
FP12B	Average number of records per month in MLF ID is 50% greater than 75th percentile – dialysis patients only	***Y***
FP12C	Average number of records per month in MLF ID is 50% greater than 75th percentile – excluding chemotherapy and dialysis	***Y***
FP13	Person ID contains at least one record with blank name and blank address	***Y***
FP14A	Overlapping non-contract hospital episodes where identifying details vary	***Y***
FP18	If sex differs between records, check sex-specific diagnoses and procedures	***Y***
FP19	Women <15 years of age at birth of last child with >4 children	***Y***
FP20A	any person ID containing >10 baby records in QPDC	***Y***
FP20B	any person ID containing >10 baby records in QHAPDC (excluding terminations)	***Y***

FP21A	Two births within 8 months in QPDC (excluding multiple births)	***Y***
FP21B	Two births within 8 months in QHAPDC (excluding multiple births and terminations)	***Y***
FP22	mother's date of birth<baby's date of birth	***Y***
FP23A	Multiple patient ID numbers at the same facility and identifying details vary	***Y***
FP27	Multiple 'Baby of' records with varying surnames	***Y***
	Checks for missed links	
FN23A	Records with matching facility id and UR number that are not identified as matches (public facilities)	***Y***
FN23B	Records with matching facility id and UR number that are not identified as matches (private facilities)	***Y***
FN24	Hospital admissions with separation mode of death without matching death registration record	***Y***
FN24A	Hospital Organ Procurement without matching death registration record	***Y***
FN25A	Hospital admissions with birth-related diagnosis code without matching birth registration record	***Y***
FN25B	QPDC baby record without matching birth registration record	***Y***
FN25C	Birth registration record without matching QPDC baby record	***Y***
FN26	Hospital admissions with birth-related diagnosis code without matching QPDC mother record	***Y***
FN27	Cross check against other linkage of same data sets (eg client directory)	
FN28	Cross check against other existing unique person or event identifiers (eg eARF for QAS data linkage)	
FN29	Check number of persons vs number of episodes over time for selected conditions	
FN30	Check readmission rate to same vs other facilities is sensible	
	Check for false positives and missed links	
OTH31	Manual clerical review of all records not identified as a match or a non-match by data matching software	**Y**
OTH32	Check of random samples against careful manual review	
OTH33	Credibility checks of the number of linkages obtained at various stages of linkage process	

Other steps taken to maximise the quality of linkage output

- Documenting data manipulation and linkage
- Checking the number of records after each step to ensure as expected
- Assessing the suitability of matching variables based on quality, known data issues for populations, and linkage engine limitations
- Cleaning and standardising matching variables
- Performing linkage separately on groups defined by data quality of matching variables:
  -standard matching for good quality variables
  -attaching additional data from other sets where available for poor quality variables such as baby names and women who may have changed their surnames
  -using alternative information such as hospital ID and dates to improve linkage where required
- Assessing the characteristics of unlinked records (quality of datasets, linkage variables, linkage strategy)

Abbreviations

ARIA	Accessibility/Remoteness Index of Australia
CHeReL	Centre for Health Record Linkage (NSW)
CIMHA	Community Integrated Mental Health Application
CSIRO	Commonwealth Scientific and Industrial Research Organisation
DG	Director-General
ED	Emergency Department
EDIS	Emergency department Information Systems
HIIRO	Health Innovation, Investment Research Office
HREC	Human Research Ethics Committee
MLF	Master Linkage File
MOU	Memorandum of Understanding
NCRIS	National Collaborative Research Infrastructure Support
NEAF	National Ethics Approval Form
PHA	Public Health Act
PHRN	Population Health Research Network
QA	Quality Assurance
QAS	Queensland Ambulance Service
QCR	Queensland Cancer Registry
QHAPDC	Queensland Hospital Admitted Patient Data Collection
QPDC	Queensland Perinatal Data Collection
REGU	Research Ethics and Governance Unit
RG	Registrar General
RLG	Research Linkage Group
SLA	Statistical Local Area
SSA	Site Specific Assessment
UQ	University of Queensland
UR	Unit Record
WADLU	Western Australia Data Linkage Unit

Glossary – definitions of key terms

Identified data	Data that allow the identification of a specific individual, either directly or indirectly (potentially identifiable), are referred to as “identified data”. Such data are deemed to be confidential.
Data Linkage	Data linkage is a process that uses person-level identifying information (such as name, date of birth) to determine which records within a data source, or between multiple data sources, pertain to a particular individual. The SSB data linkage team employs probabilistic methods to perform project-specific data linkage. Data output is provided with an anonymised person ID that the researcher may use to combine information from different sources, if applicable.
Direct Identifier	A direct identifier is information that establishes the identity of an individual or organisation. Examples of direct identifiers are name, address, driver’s licence number, patient UR number and Medicare number.
Indirect identification and potentially identifiable data	Indirect identification occurs when the identity of an individual or organisation is disclosed, not through the use of direct identifiers, but through a combination of unique characteristics. Data containing variables other than name and address can still be potentially identifiable if indirect identification of an individual or organisation is possible. For example data including a combination of date of birth, age, sex, Indigenous status and postcode may be identifiable if this combination of variables can uniquely identify an individual. In particularly small data sets, even a single variable such as a postcode may be an indirect identifier. These data are also deemed to be confidential.
Anonymised data	Data that have had any identifiers removed and would not permit indirect identification of an individual or organisation are referred to as “anonymised” or unidentifiable data and are not considered to be confidential. It should be recognised that the term “de-identified” is used frequently in documents other than this Statement to describe sets of data from which only names or other direct identifiers have been removed, i.e. ‘de-identified’ is intended to have the same meaning as ‘anonymised’ as per the definition above. In these instances the use of the term ‘de-identified’ is incorrect as these data still remain “potentially identifiable” and confidential.
Master Linkage File	A reference file containing the association between linkage keys and record identifiers from linked datasets.
Linkage Key	The codes created and stored in a Master Linkage File which can be used to group records that refer to the same entity.