Queensland Data Linkage Framework

January 2021

Queensland Government

Queensland Data Linkage Framework

Published by the State of Queensland (Queensland Health), December 2020

This document is licensed under a Creative Commons Attribution 3.0 Australia licence. To view a copy of this licence, visit creativecommons.org/licenses/by/3.0/au

© State of Queensland (Queensland Health) **2020**

Updated June 2023

You are free to copy, communicate and adapt the work, as long as you attribute the State of Queensland (Queensland Health).

For more information contact:
Statistical Analysis and Linkage Unit, Statistical Services Branch, Department of Health, GPO Box 48, Brisbane QLD 4001, email DLQ@health.qld.gov.au, phone (07)3234 0128.

An electronic version of this document is available at
https://www.health.qld.gov.au/__data/assets/pdf_file/0020/1211483/qlddatalinkageframework.pdf

Disclaimer:
The content presented in this publication is distributed by the Queensland Government as an information source only. The State of Queensland makes no statements, representations or warranties about the accuracy, completeness or reliability of any information contained in this publication. The State of Queensland disclaims all responsibility and all liability (including without limitation for liability in negligence) for all expenses, losses, damages and costs you might incur as a result of the information being inaccurate or incomplete in any way, and for any reason reliance was placed on such information.

Contents

Introduction

Purpose of this document

This document details the principles, structure and processes of Data Linkage Queensland (DLQ) which is located within the Statistical Services Branch at the Queensland Department of Health.

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 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.

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 stakeholders 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 Statistical Analysis and Linkage Unit (also known as Data Linkage Queensland, DLQ) in the Statistical Services Branch 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. Funding under the NCRIS/PHRN and from Queensland Health currently subsidises data linkage work conducted within DLQ 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.

DLQ is an accredited Integrating Authority. Integrating Authorities undertake high risk data integration projects involving Commonwealth data for statistical and research

purposes and DLQ is authorised to receive and/or link Commonwealth data with the approval of the relevant Commonwealth Data Custodian. Accredited Integrating Authorities are listed on the Australian Government website.

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 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 conducting large scale longitudinal research studies/or clinical trials or to collecting extra data to supplement an existing database when that data exists elsewhere.
- Enabling the re-use of existing data sets and adding 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.

Analysis of linked data has 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, patient satisfaction and understanding health services use and outcomes for persons diagnosed with COVID-19.

Methods of linking

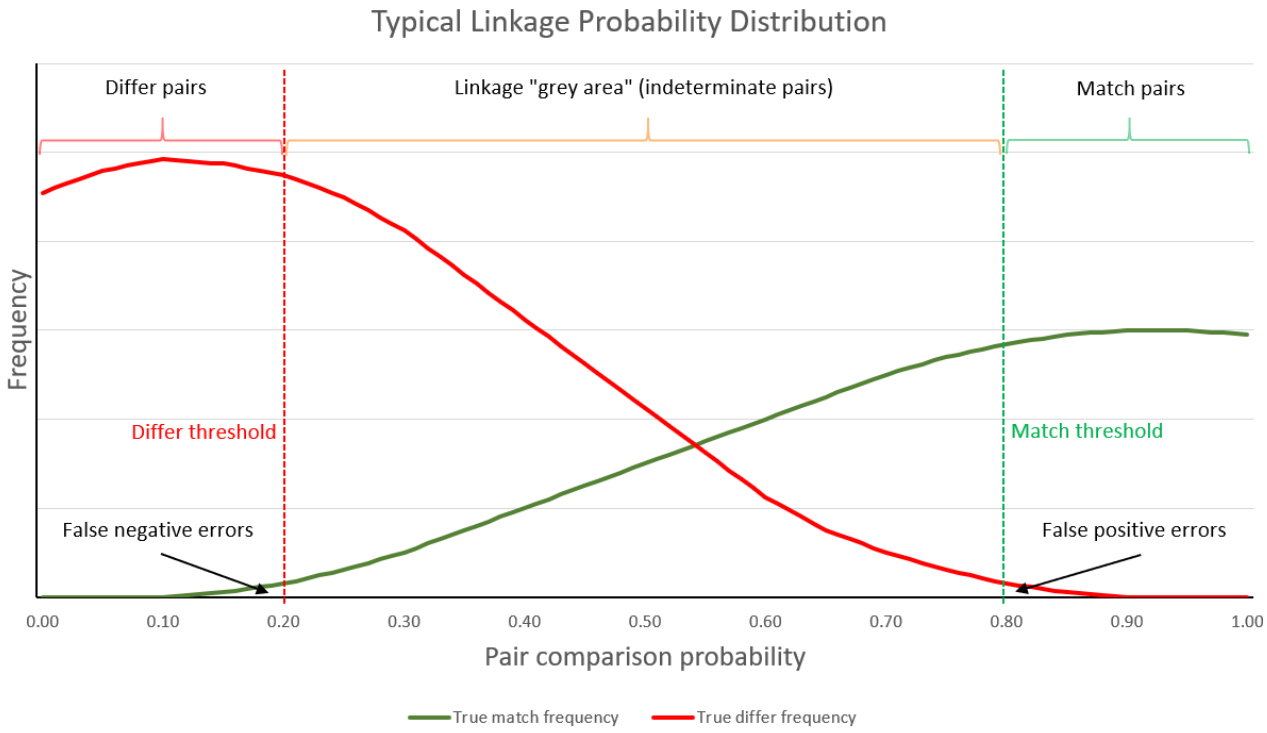
In Queensland, both deterministic and probabilistic methods of linking records are used. Both methods have their own advantages and their usage is dependent on the information available.

Deterministic linkage involves the linking of data sets using unique identifiers such as a patient/client unique identifier or through comparing fields such as name, street name, year of birth, street number with the requirement that the records agree on all characters. Deterministic linking (sometimes called exact matching) can result in missed matches when there are inconsistencies in the way information is recorded across data sources or introduce false positive matches if limited data are used to merge records. The use of computer programs and partial identifiers such as postcodes can increase the proportion of true matches to alleviate these limitations.

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 where there is less certainty about whether or not they are true matches, and human judgement is required to decide whether to link them.

A typical linkage probability distribution is shown below.



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*. DLQ follows an approval process for linkage projects consistent with the Departmental process for the release of confidential health information. Furthermore, SSB requests that approved research applicants sign the SSB Conditions of Disclosure before data is released. 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. The process and requirements for requests for research purposes 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. Further information about the process for applying for linkage services and linked data for non-research requests is available on the Statistical Services Branch (SSB) website.

Data linkage in Queensland

Datasets available for linkage

Datasets commonly included in linkage requests include:

- Queensland Hospital Admitted Patient Data Collection (QHAPDC)
- Queensland Perinatal Data Collection (QPDC)
- Queensland Cancer Registry (QCR)
- Death Registration data[1]
- Emergency Department Collection (EDC)
- National Hospital Cost Data Collection (NHCDC)
- Community Integrated Mental Health Application (CIMHA)
- Queensland Ambulance Service (QAS) data

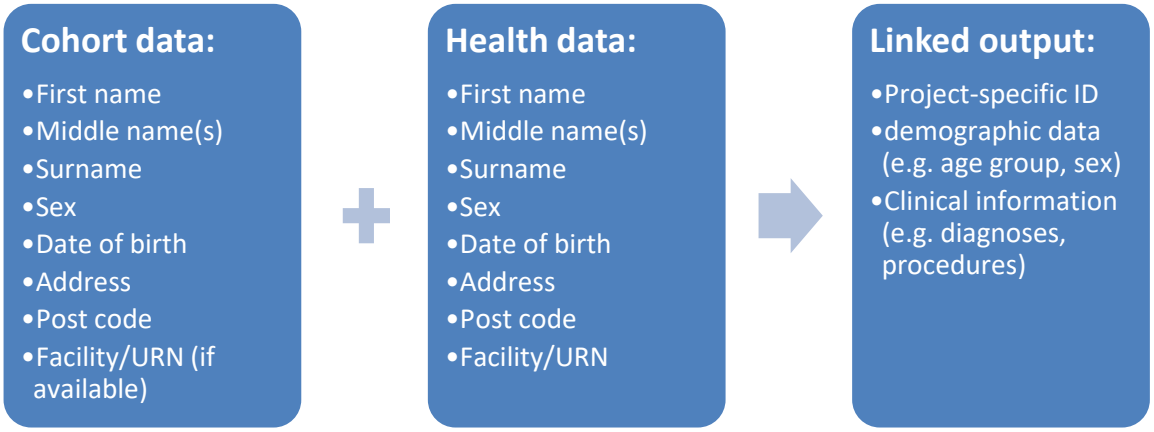
Further details about datasets are available on the SSB website.

DLQ is responsible for the ongoing production of the Queensland Health Master Linkage File (MLF) containing permanently linked references to routinely linked data collections. The use of this file saves a significant amount of DLQ resources and results in a faster processing time, benefiting researchers. The MLF is updated in near-real time such that, where feasible, data are extracted from sources as often as 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 the SSB website for details of data collections currently included in the Queensland Health MLF and for approximate lag times for submission of data for each data collection.

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

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 a file containing only information relevant to the research proposal:



In general, steps involved in supplying data for linkage and the linkage process include:

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) such as heart failure patients identified in admitted hospital activity data and their associated emergency department presentations and Qld Ambulance Service activity.
2. Client obtains all required approvals for the project (as described in the below section 'Protocol for researchers requesting data'). Once approval from the QH Director General has been obtained, the researcher is responsible for providing the signed approval documents to the SSB Client Services team and coordinating the provision of any necessary data to SSB. Linkage work commences only when all datasets have been received and linkage resources are available.
3. Where applicable, cohort file and any other non-SSB datasets containing only identifiers for linkage, plus a person and (if applicable) record ID, sent securely to DLQ. 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.
4. Linkage staff use the most efficient/appropriate method of linking cohort to required dataset(s). For example, if the MLF can be used for a linkage request this will greatly reduce the amount of clerical checking of uncertain 'grey area' links that is required. Similarly, for linkages involving QHAPDC, a deterministic matching exercise can be done if compatible identifying reference numbers are available in the cohort data, with any remaining unmatched records in the cohort linked probabilistically. A project-specific linkage key is created during this process.

5. The whole linkage process is quality-checked by a separate member of the 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 as csv files in long form.
7. Final output and report are sent to the client via a secure method of transmission. Normally, Kiteworks is used to securely transfer data. If the client is not able to receive data in this way, an alternative secure transfer 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 types of Quality Assurance checks applied to linked data in Queensland is shown in Appendix 2. These checks are applied as an additional layer of verification following matching by linkage software and manual review of all grey area identified.

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

Cohort data

- Check if cohort dates (e.g. date of birth, admission/procedure dates) are within the range specified in the request documentation.
- Identify and quantify missing values in identifying variables.
- Check for duplicate records if the cohort file should consist of distinct entities.
- Clean and standardise identifying variables to optimise consistent matching with other 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 inform whether or not they are to be considered to be 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, could these be twins?
- If date of birth does not match, does it appear to be estimated in one dataset? For example, a commonly used value where the actual date is not known is 1 January of the same or an adjacent year.
- Could a difference in surname be due to marriage or misspelling? Although less common than in females, there are instances where males (particularly children) have changed surname. Where it is suspected that a name change may have occurred these are checked by the data linkage team against marriages and name change data held by the Registry of Births Deaths and Marriages via a secure online service.
- Address changes are common, but geographical distance can be considered in deciding if two records with different addresses relate to the same entity.

In addition, the proportion of linked to non-linked cohort records will be considered. For example where 100% of records in a cohort are expected to be present in a health data collection but not all records are linked, unlinked cohort records will be examined for characteristics such as missing/poor quality information and a manual search may be undertaken in the relevant health datasets to establish a link.

Final output

Output is checked for issues such as:

- Instances where date of death precedes other dates, such as hospital admission.
- Compliance with logic that is applicable to a project, such as one distinct cohort person linking to each health record.
- Oddities such as a woman apparently giving birth twice within six months.
- Unrealistic and out-of-range values of variables like age.
- Multiple death records linking to one person.

Limitations and role of linkage clients in quality assurance

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. Errors also occur within data collections which may result in quality errors that can impact on analyses.

Analysts working with linked output should apply their own quality checks and report any potential linkage quality issues back to DLQ to support ongoing quality improvements. The New South Wales data linkage unit, Centre for Health Record Linkage (CHeReL), has published useful guidance about data quality checks.

Protocol for researchers requesting data

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 2005*. The unit that coordinates research applications within Queensland Health is the Office of Research and Innovation (ORI).

Office of Research and Innovation

ORI, previously known as the Health Innovation, Investment and Research Office (HIIRO), is the central contact point for advice about the process for applying for Queensland Health data for research purposes under the *Public Health Act 2005* and for information about ethical and governance issues associated with the conduct of research using Queensland Health data.

Research requests for Queensland Health data must have clearance from an ethics and governance perspective and from a legal perspective. Details of these requirements are described below.

When all required approvals have been obtained for a project and provided to ORI, ORI will assess the Public Health Act application (PHA) and, if appropriate, issue an approval notice from the Queensland Health Director-General (D-G) or delegate for final approval. Once this final approval is granted, ORI will record the approval in the Research Registry and send a letter to the researcher advising of final approval. The researcher then needs to provide a copy of this notice from ORI to data custodians (in the case of SSB, email to DLQ@health.qld.gov.au). Once the D-G or delegate approval for the data release has been received in SSB, the request will be placed in a queue awaiting linkage resources.

The flowchart below shows the steps involved in applying for confidential Queensland Health data for research purposes.

Ethics and Governance requirements

Requirements include:

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). The ethics approval must cover the entire period where data are being analysed. An extension or amendment must be obtained if data are to be used outside of the period covered by the initial ethics approval.

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 ORI's website or contact ORI for more information.

Legal requirements

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 application (PHA). This application requests information about 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 and requires data custodian sign-off. The Director-General (DG) or their delegate may grant access to health information for the purpose of research based on this application if it is in the public interest.

The PHA form requires the signature of each data custodian in order to be further processed by ORI. It is important that all Data Custodians involved in the project review and approve the same version of the PHA. Once the Data Custodian has signed the PHA form at section 10, the PHA will be scanned and emailed as a PDF back to the researcher who then needs to forward all documents to ORI.

The PHA form 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, such as requests for additional data items, the PHA form must be re-submitted to the relevant Data Custodian(s) and go through the entire approval process again.

Please refer to ORI's website for further information. Queries about the application process for research projects that involve data linkage can be emailed to DLQ@health.qld.gov.au.

Obtaining Data Custodian approval on Public Health Act applications

The Data Custodian 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 form. The PHA requires Data Custodian approval for each dataset listed in the application. This may result in approval from multiple Data Custodians.

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 planning on accessing data linkage services should also contact the data linkage client services team on DLQ@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. It is recommended that researchers contact the data linkage client services team before seeking approval from Data Custodians so that linkage details can be included in the draft PHA.

SSB is the current Data Custodian for QHAPDC, non-admitted patient data, QPDC and birth and death registration data, as well as the provider of linkage services.

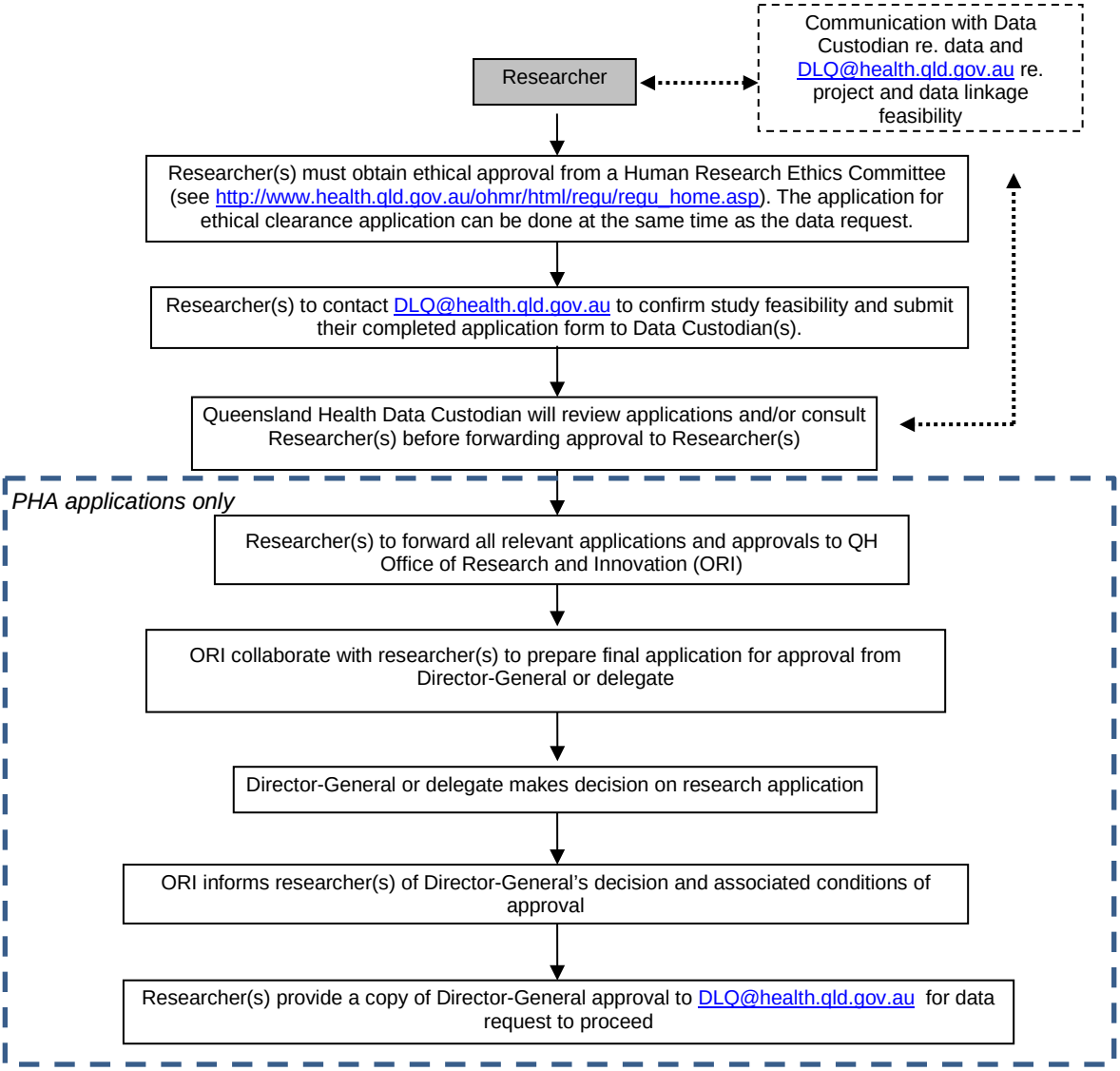
The Data Custodian has the responsibility of ensuring that appropriate patient privacy is maintained when considering requests for data. For this reason, only those Queensland Health data that are clearly shown to be essential for the research project are considered for release.

ORI maintains a list of Data Custodians for Queensland Health data collections. Additionally, SSB has resources about data collections commonly requested for linkage including Data Custodian details and links to information about data collections.

Signed Consent

If a researcher is able to obtain signed consent from participants to use their personal or identifying information for a clearly specified research study, then a Public Health Act application may not be required for health information held by Queensland Health, as the disclosure of information is authorised by another act (for example, Section 144 of the *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 Data Custodian also has the discretion to request a copy of signed consent forms.

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



Penalties for misuse of linked data

The Australian Code for the Responsible Conduct of Research (2018) 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. Linked data must only be used for the research purposes outlined in the ethics and approved application.

From 28 September 2016 changes to the *Privacy Act 1988* made it illegal to re-identify government data which has been "de-identified".

In addition, to comply with national privacy legislation (National Privacy Principle 9) data can only be accessed by approved researchers from locations within Australia. Data provided by Queensland Health are not to be accessed from overseas locations or stored on servers located overseas.

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 should be referred to when trying to determine what level of information to request.

Some points to consider when negotiating with Data Custodians or submitting an application:

- The scope of the data (cohort of interest, variables and dates requested, etc.) needs to be consistent with and justified by the purpose of the research. Applicants must specify the inclusion/exclusion criteria for the study group of interest and requested dataset. For example, if data are only required for acute episodes for children under 10 years of age who reside in a specific area, then these criteria must be clearly stated in the data request. Broad data requests (e.g. asking for all episodes of care for all ages) are not likely to be approved unless clearly justified by the research questions being investigated.
- 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 despite the research project being focussed on a specific health condition of interest.
- Check whether the data items or named information are available for the entire period requested. For example, QPDC mother names and QHAPDC private facility patient names were not collected until July 2007 so data linkage is not possible prior to this time (see the Master Linkage File page on the SSB website for further details regarding data availability).
- Ensure the conditions and/or procedures of interest are clearly defined. The relevant ICD diagnosis or ACHI procedure codes for the period of interest should be stated in the application.
- The 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 years of age;
    - 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/SA2;
- 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 secure, encrypted data transfer method such as Kiteworks with the password communicated by phone on receipt of data is preferred.
- Review how the data will be stored. For example, a secure university network server may be a viable option. The data custodian should be assured that confidential patient information will be held securely and accessible by only those researchers named in the application.
- Data output is generally provided in relational tables.
- 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

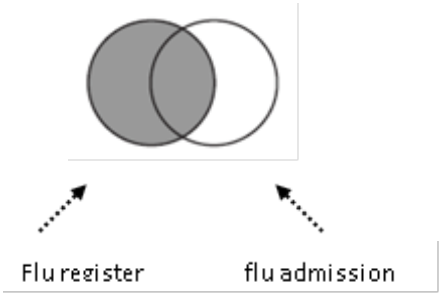
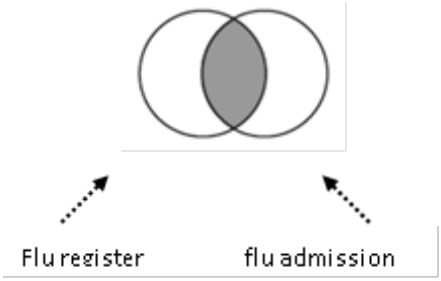


Figure 2



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 with hospitalisation data for a subset of this group, 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.

Appendices

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

Resources linked in this document

Statistical Services Branch (SSB): https://www.health.qld.gov.au/hsu

Data linkage in Queensland (SSB): https://www.health.qld.gov.au/hsu/link/datalink

How to access data from SSB: https://www.health.qld.gov.au/hsu/how-to-access-data-from-ssb

Data collections in Master Linkage File (SSB):
https://www.health.qld.gov.au/hsu/link/datasets

Office of Research and Innovation (ORI):
https://www.health.qld.gov.au/hiiro
https://www.health.qld.gov.au/hiiro/html/contact_us

Public Health Act research requests (ORI):
https://www.health.qld.gov.au/hiiro/html/regu/aces_conf_hth_info

Information for researchers and sponsors (ORI):
https://www.health.qld.gov.au/hiiro/html/regu/for_researcher

Data file format and transfer (SSB):
https://www.health.qld.gov.au/__data/assets/pdf_file/0036/881685/dataformat.pdf

Data file passwords and encryption (SSB):
https://www.health.qld.gov.au/hsu/link/training/passwords-and-encryption

Data format, coding and efficiency (SSB):
https://www.health.qld.gov.au/hsu/link/training/format-coding-and-efficiency

Edit checks before analysing linked data (CHeReL):
https://www.cherel.org.au/media/24762/edit_checks_-_Oct12.pdf

Public Health Act application amendments (SSB):
https://www.health.qld.gov.au/__data/assets/pdf_file/0031/818095/updatedpha-ssb.pdf

Queensland Health data custodian list (ORI):
https://www.health.qld.gov.au/__data/assets/pdf_file/0034/843199/data_custodian_list.pdf

Commonly requested data collections (SSB): https://www.health.qld.gov.au/hsu/crdi

Research requests using consent (SSB): https://www.health.qld.gov.au/hsu/consent

Australian Code for the Responsible Conduct of Research (NHMRC):
https://www.nhmrc.gov.au/about-us/publications/australian-code-responsible-conduct-research-2018

Guidelines for the Use and Disclosure of Health Data for Statistical Purposes:
https://www.health.qld.gov.au/__data/assets/pdf_file/0022/374701/simc_anonymisation.pdf

Potentially identifiable data (SSB): https://www.health.qld.gov.au/hsu/potentially-identifiable-data

Relational data (SSB): https://www.health.qld.gov.au/hsu/link/training/relational-data

Other useful resources

SSB data collections: https://www.health.qld.gov.au/hsu/collections/dchome

Commonly requested data items: https://www.health.qld.gov.au/hsu/crdi

Requesting data from SSB: https://www.health.qld.gov.au/hsu/how-to-access-data-from-ssb
Data linkage training resources: https://www.health.qld.gov.au/hsu/link/training/data-linkage-training-resources

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)

Appendix 2 – Examples of quality checks performed by the Queensland Research Linkage Group

Error ID	Checks for false positive links	Scheduled[2] 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
FP06D	Person ID contains multiple admitted patient data death records	Y
FP07	A woman is unlikely to have been < 13 when she first gave birth	Y
FP07A	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 (mother)	Y
FP07B	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 (baby)	Y
FP08	A woman is unlikely to have been > 50 when she last gave birth	Y
FP08A	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 (mother)	Y
FP08B	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 (baby)	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	Very high avg monthly events are less likely to be 1 person	Y
FP12A	Person ID having more than 15 episodes in a month - chemotherapy patients only	Y
FP12B	Person ID having more than 15 episodes in a month - dialysis patients only	Y
FP12C	Person ID having more than 15 episodes in a month - excluding chemotherapy and dialysis	Y
FP12D	Person ID having more than 15 episodes in a month - EDC	Y
FP13	Person ID contains at least one record with blank name and blank address	Y

[2] Checks are performed at various intervals, ranging from monthly to triennially.

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
FN23A	Records with matching facility id and UR number that are not identified as match (public)	Y
FN23B	Records with matching facility id and UR number that are not identified as match (private)	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
FN24B	Death record with hospital ID populated, without any in-hospital death or any other record at that facility	Y
FN24C	Emergency data with discharge status of death 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
FP27	Multiple ‘Baby of’ records with varying surnames	Y
FP28	Previously separated linkage keys automatically merged (by ChoiceMaker)	Y
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	Reasonability checks of the number of linkages obtained at various stages of the linkage process	
OTH34	Cross check against other linkage of same data sets (eg client directory)	
OTH35	Cross check against other existing unique person or event identifiers (eg eARF for QAS data linkage)	
OTH36	Check number of persons vs number of episodes over time for selected conditions	
OTH37	Check readmission rate to same vs other facilities is sensible	

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)
- Perform inter-rater assessments and random sampling to ascertain and improve consistency of clerical reviewer decision-making
- Correctively action reported MLF errors

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
DLQ	Data Linkage Queensland
ED	Emergency Department
EDIS	Emergency department Information Systems
ORI	Office of Research and Innovation
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 application
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
SA2	Statistical Area Level 2
SSA	Site Specific Assessment
UQ	University of Queensland
UR	Unit Record

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. ‘De-identified’ is not the same 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.