eHealth Queensland

De-Identification and Anonymisation of Data Guideline

V1.0

Queensland Government

Health Informatics Services

Published by the State of Queensland (Queensland Health), March 2021

Security classification	IN-CONFIDENCE		
Document licence	N/A		
Copyright	© State of Queensland (Queensland Health) 2021		
Authority	State of Queensland (Queensland Health)		
Author	eHealth Queensland		
Documentation status	Final	Release	1.0

Further information about security classifications is available in the

Information security classification framework (QGISCF)

For more information contact:
Health Informatics Services, Digital Strategy and Transformation Branch, e-Health Queensland, email eHealth-IMStrategy@health.qld.gov.au

Version history

Version	Date	Status	Key changes made	Author/s
1.0	16 March 2021	Final	Approved by the Information Management Strategic Governance Committee (IMSGC)	Health Informatics Services

Contact for enquiries and proposed changes

If you have any questions regarding this document or if you have a suggestion for improvement, please contact:

Contact Officer Health Informatics Services,
Digital Strategy and Transformation Branch,
eHealth Queensland

Email eHealth-IMstrategy@health.qld.gov.au

Table of Contents

1. Introduction

1.1 Purpose

The purpose of this document is to provide guidance and direction when de-identifying and/or anonymising personal and/or confidential and/or sensitive information for the purposes of authorised use and disclosure. This document provides information to support the process of de-identifying and/or anonymising data and outlines available techniques used to support the process. This document addresses the data minimisation principle and offers advice about, rather than stipulating which, technique to use for each given scenario.

The purpose of this document is not to define the use cases for which de-identification and/or anonymisation of data may need to occur. Similarly, the purpose is not to define the end to end considerations required when de-identifying and/or anonymising data. Section 3 clarifies the scope of this document.

It is understood that in the event that a contract is being formed, and a definition of 'de-identified data' is required, the relevant legal advisors will need to be consulted. Definitions contained in legal documents are not a 'one size-fits-all', therefore a legal review of the specific circumstances will be required. The definitions will not be able to be used in contractual documents without the approval by the relevant legal advisors.

1.2 Audience

The intended audience for this document is:

- Relevant Queensland Health stakeholders involved in the disclosure and use of data.
- Relevant Queensland Health stakeholders involved in the de-identification and/or anonymisation of data.

2. Background

2.1 Context

There is an increasing recognition of the value of sharing and releasing data enabling it to be analysed and used, for example, for research, to improve health planning and for predictive analytics to target patient care or for monitoring disease outcomes.

All Queensland government agencies deal with personal information. In doing so, they must comply with the privacy principles in the *Information Privacy Act 2009* (Qld) (IP Act). Health agencies are required to comply with the privacy principles including the nine National Privacy Principles set out in the IP Act. NPP 2 provides that personal information may only be used for the purpose for which it was obtained and not for any other purposes, unless one of the exceptions applies. Where information has been appropriately de-identified, it is no longer personal information and can therefore be used or shared in ways that may not otherwise be permitted under the *Information Privacy Act 2009* (Qld).[1] For example, de-identified data from the My Health Record (MHR) system could be safely integrated with de-identified data from other government systems for public benefit. De-identified MHR data, immigration data and hospital admissions data, in a specified city, could be safely linked to determine whether long-haul flights contribute to higher instances of deep-vein thrombosis.[2] Anonymised data, collected from Hospital and Health Services relating to admissions each year, could enable government departments to understand the catchment area for each facility, the admission diagnoses, number of admissions for each patient as well as length of stay and their age group. Information could be analysed and used to determine whether the current number of facilities adequately provide the services required.[3]

The privacy principles in the IP Act operate subject to other Acts that deal with the disclosure of information. This includes the *Hospital and Health Boards Act 2011* (Qld) (HHB Act) Pt. 7 which applies to confidentiality of patient information. Health agencies cannot rely on NPP 2 to disclose information where it is prohibited by the HHB Act, or by another piece of legislation. Both part 7 of the HHB Act and the IP Act do not apply to de-identified information or statistical datasets that do not allow individuals to be identified.

De-identification and anonymisation are processes which support the sharing or dissemination of data ethically and legally, thereby realising its social, environmental and economic value, whilst preserving confidentiality. De-identification and anonymisation are used by agencies for the protection of confidential and sensitive information, to build trust and meet community expectations around the handling of data.[4] However, de-identification can be technically complex and often requires specialist advice. De-identified data is also at risk of re-identification. This often occurs when de-identified data is linked with other external information. Re-identification can reveal personal information and may breach the privacy principles. When agencies release de-identified data, they must adequately manage re-identification risk to protect the identity of individuals and their personal information.

Privacy and confidentiality legislation, Memoranda of Understanding (MOUs) and other agreements set the standards for how Queensland public health agencies handle personal information[5] and confidential

[1] Office of the Australian Information Commissioner, De-identification and the Privacy Act (March 2018)
[2] Framework to guide the secondary use of My Health Record system data
[3] AIHW Guidelines for the Disclosure of Secondary Use Health Information for Statistical Reporting, Research and Analysis
[4] Office of the Australian Information Commissioner, De-identification and the Privacy Act (March 2018)
[5] Personal information is information or an opinion, including information or an opinion forming part of a database, whether true or not, and whether recorded in a material form or not, about an individual whose identify is apparent, or can reasonably be ascertained, from the information or opinion.

information including rules about collection, storage, use and disclosure of personal and confidential information.

Please note: From an open data perspective, custodians and all state government agencies who publish to the open data portal, are referred to the Office of the Information Commissioner (OIC) Queensland guidelines.

3. Scope

Figure 1 describes the phases of de-identification or anonymisation of data that are in scope and out of scope for this document.

The in scope and out of scope phases have been determined based on ability to address the content adequately within one document, not to underestimate the importance of all the phases. It is expected that over time, guidance documentation will be developed to support all phases.

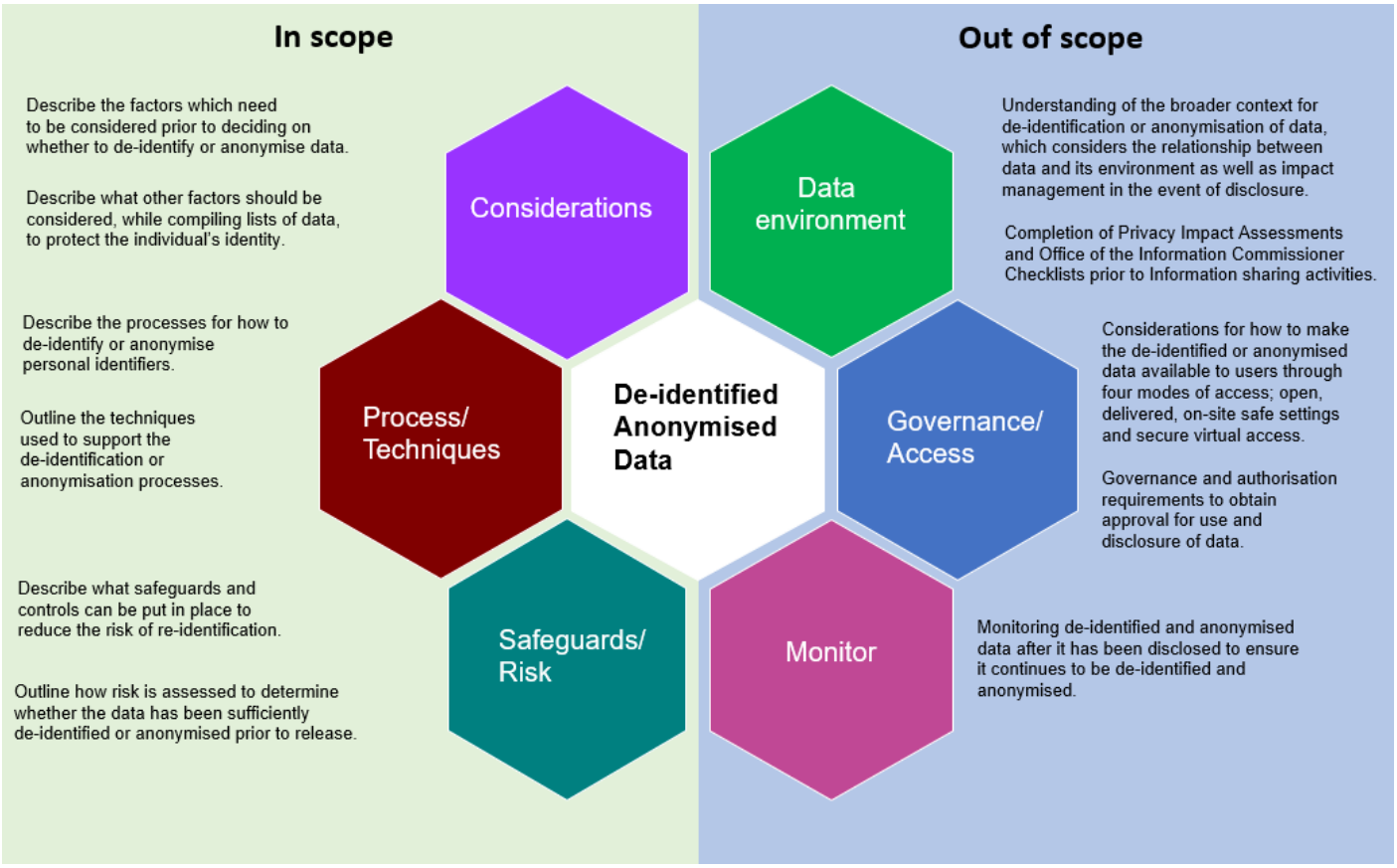

Figure 1: Phases of the de-identification or anonymisation of data that are in scope and out of scope for this document

For further information, please refer to the CSIRO, The De-Identification Decision-Making Framework (Sep. 2017)
For a guide to handling security breaches, refer to OAIC, Data breach preparation and response.

4. Assumptions

- Specialised skills and knowledge are required to correctly de-identify or anonymise a dataset to minimise the risk of identification. Quality assurance should also be performed to verify that data has been correctly de-identified or anonymised.

5. De-identification vs Anonymisation of data: A Comparison

The terms de-identification and anonymisation of data are often used interchangeably which can cause confusion. To avoid any misunderstanding, it is important to be aware of the differences and nuances in the context of how these terms are used.

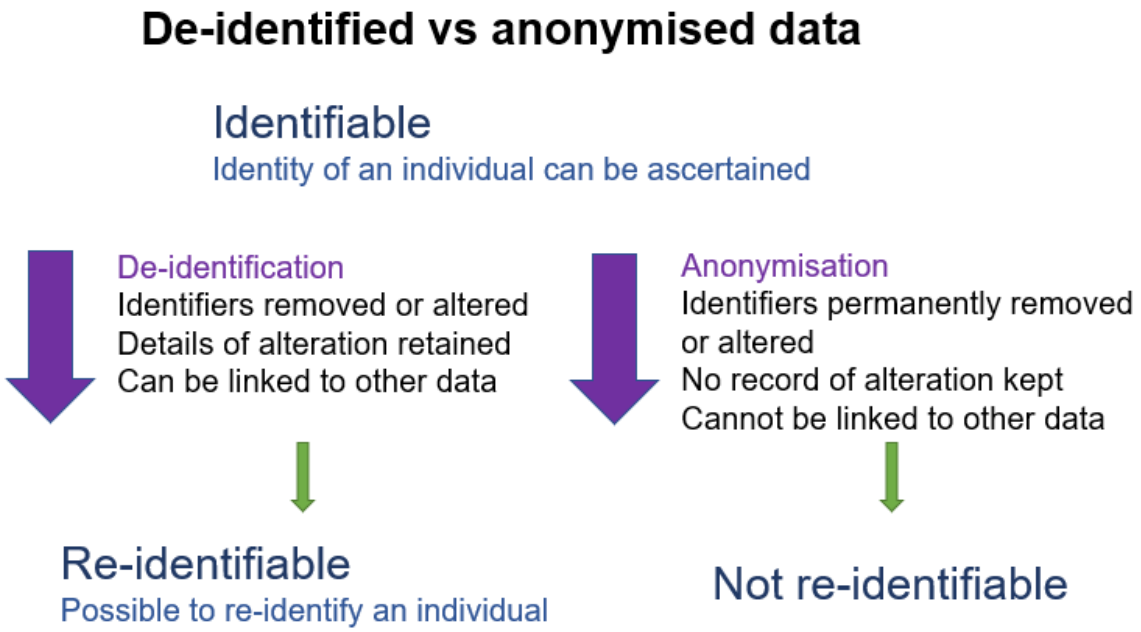

Figure 2: Comparison of de-identification and anonymisation of data

Table 1 provides a comparison between de-identification and anonymisation of data which is recorded against each attribute of the process.

TABLE 1. DATA DE-IDENTIFICATION VS DATA ANONYMISATION

Attribute it relates to	De-identification	Anonymisation
Process	A process which involves the removal or alteration of personal identifiers, followed by the application of any additional techniques or controls required to remove, obscure, aggregate, alter and/or protect data in some way so that it is no longer about an identifiable individual.[6]	The process of ensuring that the risk of somebody being identified in the data is negligible. This invariably involves doing more than simply de-identifying the data, and often requires that data be further altered or masked in some way in order to prevent statistical linkage.[7]
HHB Act Interpretation	If the information relates to 'a person who is receiving or has received a public sector health service' where the information could identify the person, the information is still confidential and subject to the constraints of the HHB act, Part 7.	No longer considered confidential information.
IP Act	If the data is no longer about a person 'who is reasonably identifiable' the data is not considered personal information.	No longer considered personal information.
Identification	The term de-identified is used to describe data that is no longer about an identifiable individual and does not reveal personal information about such an individual.[8]	Wherever it is lawful and practicable, individuals must have the option of not identifying themselves when entering into transactions with a health agency.[9]
Re-identification	De-identification is the removal of identifying information from a dataset, and this data could potentially be re-identified, e.g. if the identifying information is kept (as a key) and recombined with the de-identified dataset.	Anonymisation is the permanent removal of identifying information, with no retention of the identifying information separately.[10]
Risk	De-identification techniques are applied, using varying levels of alteration, depending on the particular use of the data and type of environment it is being released into.	Anonymisation is about reducing the risk of privacy breach to a negligible level.

[6] OAIC, De-identification and the Privacy Act
[7] The Anonymisation Decision-Making Framework
[8] CSIRO, The De-Identification Decision-Making Framework (September 2017)
[9] *Information Privacy Act 2009* (Qld)
[10] Australian National Data Service Guide

6. De-identify data

6.1 Definition

De-identification is a process involving the removal or replacing of direct identifiers in a dataset, followed by the application of any additional techniques or controls required to remove, obscure, aggregate, alter and/or protect data in some way so that it is no longer about an identifiable individual. This will usually require that the risk of other types of disclosure, such as attribute disclosure or inferential disclosure, are very low.[11]

6.2 Considerations

Prior to de-identifying data, the following three (3) core activities need to be conducted: (1) data situation audit, (2) risk analysis and control, and (3) impact management.

It is important to remember that de-identification is not a fixed or an end-state. Like other risks, re-identification risks and their controls require ongoing monitoring and review. The risk of re-identification increases as technology develops and/or as more 'auxiliary information' is published or obtained by a person or entity.

The same information may be personal information in one situation but de-identified information in another. For example, a custodian of the data de-identifies data but retains a copy of the original dataset. This may enable them to re-identify the data subjects in the de-identified dataset. So, the dataset may be personal information when handled by the custodian of the data but, may be de-identified when handled by a different authorised entity because the data access environment is different.[12]

De-identified data could potentially be re-identified. Re-identification could occur when data is combined with external 'auxiliary information' to reveal information about an individual. A re-identification event can reveal:

- Personal information and may breach the privacy principles in the *Information Privacy Act 2009* (Qld).
- Confidential information (i.e. patient information) and may breach the duty of confidentiality in Part 7 of the *Hospital and Health Boards Act 2011* (Qld). [13]

The Five Safes framework is an approach to thinking about, assessing and managing risks associated with data sharing and release, and comprises five dimensions.

Table 2 lists the Five Safes' dimensions, their meanings and how to interpret them.

TABLE 2. THE FIVE SAFES' DIMENSIONS

Dimension	Meaning	Interpretation
1. Safe projects	Is the use of the data appropriate?	Use of the data is legal, ethical and the project is expected to deliver public benefit.

[11] CSIRO, The De-Identification Decision-Making Framework (September 2017)
[12] Office of the Australian Information Commissioner, De-identification and the Privacy Act (March 2018)
[13] Confidentiality General Principles *Hospital and Health Boards Act (HHB) 2011*

Dimension	Meaning	Interpretation
2. Safe people	Can the users be trusted to use it in an appropriate manner?	Researchers have the knowledge, skills and incentives to act in accordance with required standards of behaviour.
3. Safe data	Is there a disclosure risk in the data itself?	Data has been treated appropriately to minimise the potential for identification of individuals or organisations.
4. Safe settings	Does the access facility prevent unauthorised use?	There are practical controls on the way the data is accessed – both from a technology perspective and considering the physical environment.
5. Safe output	Are the statistical results non-disclosive?	A final check can be required to minimise risk when releasing the findings of the project.[14]

A data situation involves the relationship between data and its environment, where the environment comprises people, other data, infrastructure, and governance structures. Data releases need to be appropriate to their release environment and avoid risk of disclosure. There is the risk that indirect identifiers in combination with auxiliary information might identify an individual. The risk of potentially identifying data can be lessened by modifying the data, e.g. using aggregation where, depending on the use case, detail may be minimised accordingly on the key variables to reduce the measurable risk. When the data situation is sensitive the decision needs to be made whether to remove or reduce detail on sensitive variables.

The risk of potentially identifying data can also be lessened by reconfiguring the environment, which essentially involves controlling who has access, how they access the data and for what purposes.

A register should be kept of all de-identified data, which has been shared or released, to take account of the possibility of linkage between releases leading to a disclosure. Changes to the data environment should be monitored as they may impact the de-identified data presenting the possibility of re-identification.[15] In more extreme cases, agencies may consider removing or restricting access to the data. Details of de-identification alteration which have been recorded, for possible re-identification or data integrity checks, should be securely protected against unauthorised access.

6.2.1 Non-text based data

Images

De-identifying and anonymising medical images can present problems, especially when dealing with Digital Imaging and Communications in Medicine (DICOM) image files commonly used for computed tomography scans (CTs), magnetic resonance imaging (MRI) and positron emission tomography (PET). Individual patient scans may have several linked files that contain additional information associated with the image (i.e. contours drawn on the image and information regarding how two image formats taken during the one scanning session have been registered). Completely anonymising DICOM image files can sometimes mean the image and any additional information cannot be imported, linked or viewed in some software.

[14] AIHW Data Governance Framework
[15] CSIRO The De-Identification Decision-Making Framework

Some information on medical images may not be overwritten, when specified by vendors however, these images cannot be accessed or easily accessed in most image viewing software, meaning the images are de-identified. However, re-identification can occur in specific scenarios therefore de-identified images should be tested to determine exactly what can be viewed in the software where images will be normally viewed /analysed in. Once this has been evaluated, access to the software should be restricted, e.g. to only researchers with password protected access to the software or computer the software is on. Restrictions should also be placed on how the images will be used or presented.

Biological samples

Other datasets that may contain potentially-sensitive information include epidemiological surveys of health, medical trial data, biological sampling and genetics, which require modification of some aspects of the dataset to protect the individual's identity.

For further information please refer to Australian National Data Service Guide - Publishing and sharing sensitive data.

6.3 Process

The de-identification process involves two (2) steps: (1) removal of direct identifiers (such as name, address, driver licence number, telephone number, photograph or biometrics), see: Appendix C: Direct identifiers and (2) taking one or both of the following additional steps:

- the removal or alteration of other information that could potentially be used to re-identify an individual (such as date of birth, gender, profession, ethnic origin, marital status etc), and/or
- the use of controls and safeguards in the data access environment to prevent re-identification.[16]

Appropriately skilled individuals, e.g. custodians of data, trained clinical and/or scientific individuals, should correctly de-identify a dataset to minimise the risk of identification by:

- Applying such principles and methods to determine that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information.
- Documents the methods and results of the analysis that justify such determination.[17]

6.4 De-identification techniques

There are several techniques used to support the de-identification process. De-identification techniques should be carefully chosen (based on a risk assessment) to ensure that the individual's privacy is protected, and that the information will still be useful for its intended purpose after the de-identification process.

Before selecting a technique for de-identifying data, it is important to consider why the data needs to be shared or released and what data is appropriate, who will access the data and how those accessing the data might want to use it. Consideration of these points will help decide what data can be safely shared or released and determine the appropriate level of de-identification to be applied. For example, releasing

[16] Office of the Australian Information Commissioner, De-identification and the Privacy Act (March 2018)
[17] Guidance on De-identification of Protected Health Information (November 2012)

statistical reports to the general public would need a greater degree of de-identification to prevent disclosure than statistical reports shared between health departments.

Choosing an appropriate de-identification technique requires consideration of contextual factors such as:

- The kind of information or data that is to be de-identified.
- Who will have access to the information, and what purpose this is expected to achieve.
- Is the data request for a data linkage project.
- Whether the information contains unique or uncommon indirect or quasi-identifiers that could enable re-identification.
- Whether the information will be targeted for re-identification because of who or what it relates to.
- Whether there is other information available that could be matched up or used to re-identify the de-identified information; and
- What harm may result if the information is re-identified?[18]
- Availability of software to perform the chosen technique.

Data lies on a spectrum with multiple degrees of identifiability depending on the type of techniques applied. The spectrum ranges between explicitly personal data and potentially identifiable data to not readily identifiable data.

Explicitly identifiable data has the direct and indirect identifiers left intact and there are safeguards in place. This type of data could be shared between clinicians involved in the treatment of the same patient who has given consent for their information to be shared. Where identifiable data is used for approved research purposes, e.g. for analysis, then ethical and privacy requirements need to be met through access control and data security.

Potentially identifiable data has the direct identifiers partially masked, but the indirect identifiers are left intact and there are safeguards in place. This type of data could be shared in a scenario where a Primary Health Network (PHN) approaches a GP practice in order to obtain population health data to improve the provision of primary care support.

Not readily identifiable data has the direct identifiers partially masked, but the indirect identifiers are left intact and there are controls in place. This type of data could be used for training clinical staff purposes.[19]

[18] Office of the Australian Information Commissioner: De-identification and the Privacy Act, March 2018
[19] Australian National Data Service Guide

Table 3 provides an overview of the more commonly used techniques (see Appendix D for additional de-identification techniques) used to de-identify data for specific use cases, together with their associated impact on risk and data utility. Determination of which method is most appropriate for the data being released should be assessed by an expert on a case-by-case basis and be guided by input from the health care providers.

TABLE 3 COMMONLY USED DE-IDENTIFICATION TECHNIQUES

Technique	Details	Impact on risk	Impact on utility	Uses
Data reduction (protects tables by either combining categories or suppressing cells)				
Remove direct identifiers	• Remove details that identify a person. • Refer to: Appendix C: Direct identifiers	• Minimises risk of identification	• Minimal data available to answer the question.	• Research statistics
Omission of specific dates	• Specific dates should not be provided unless absolutely necessary. In most cases the request can be satisfied by using one or a combination of the following: - Provision of month and year of admission/separation etc - Provision of day of the week and time of day (generally for emergency department data) - Provision of dates encrypted as the number of days from the date of first event (day zero) or other selected starting date which is not known to the user. This is in order to enable the user to identify episode chronology and calculate intervals between events - Include sequence number and days between the sequential episodes.[20]	• Minimises risk of identification	• Enables identification of episode chronology • Enables calculations to be performed between events	• Data linkage • Unit record data releases
Sampling fraction	• Only a sample of the dataset rather than the whole dataset is released. • The sampling fraction is specified by the study design. • Need to consider the goals of the data release and how widespread the use is likely to be.	• Provides some protection against identification risks because it reduces the certainty about whether a particular individual or organisation is in the data, so increases the probability of false positive matches. • Even a 95% random sample creates uncertainty. • Cannot protect against types of re-identification where a third party matches a dataset with another overlapping dataset. • Needs to be used with other techniques.[21]	• Do not distort the data and are transparent in their effects. • Modest impact. • It will increase the variances of any estimates and reduce statistical power. • Where analysis of small sub-populations is required sampling may reduce the capacity to do this.	• Statistical data on a percentage of the population. • Surveys.
Choice of variables	• Certain variables should be excluded from the released dataset; a formal identifier or a quasi-identifier (e.g. significant dates, profession, income) that are unique to an individual, or which in combination with other information are reasonably likely to identify an individual. • Key variables should be removed, obscured or aggregated. • Target variables should be removed, obscured or aggregated. • With microdata, the choice is whether a variable appears in a dataset or not. • With aggregate data, the choices are about which variables will be included in each table. • For point to point data shares the variable selection will be driven by the requirements of the user.	• Key variables form the basis of any attack therefore, a reduction in the number of key variables will decrease the risk of identification. • The impact of variable selection on risk is dependent on the variables selected. • If key variables are de-selected the re-identification risk will be reduced. • The effect is to reduce key power; the capacity of a set of key variables to discriminate between records and produce both sample and population uniques. • If target variables are de-selected the sensitivity of the data is lessened and the potential impact of any breach reduced.	• Removal of a variable critical to a user's analytical requirements will disable the analysis. • With data releases consideration should be made for how widespread the use is likely to be and whether the goals of release can be met through a more modest variable selection. • Some loss of utility with removal or key variables. • Do not distort the data and are transparent in their effects.[22]	• Observational studies • Statistical reports

[20] AIHW Guidelines for the Disclosure of Secondary Use Health Information for Statistical Reporting, Research and Analysis (February 2019)
[21] The De-Identification Decision-Making Framework: appendices (September 2017)
[22] The De-Identification Decision-Making Framework: appendices (September 2017)

Technique	Details	Impact on risk	Impact on utility	Uses
Data modification (changes all non-zero cells by a small amount without reducing the table's overall usefulness for most purposes)				
Rounding	• Combine information or data that is likely to enable identification of an individual into categories. • E.g. age may be combined and expressed in ranges (25-35 years) rather than single years (27, 28). Extreme values above an upper limit or below a lower limit may be placed in an open-ended range such as an age value of 'less than 15 years' or 'more than 80 years'.	• Rounding can be very effective in reducing risks when considering individual tables of counts. • Need to consider the interactions between multiple outputs and particularly how to resolve the issue of additivity and consistency between marginal totals in different tables.	• For many purposes rounded frequencies are sufficient and using percentages as a form of rounding can be an even more digestible way of presenting information. • The slight alteration of small cells in a table ensure results from analysis based on the data are not significantly affected.[23]	• Rounding is a technique most commonly used with tables of counts.
Cell suppression	• Data are only partially released • Unsafe cells are suppressed and replaced by a special character, such as '..' or 'X', to indicate a suppressed value (primary suppression). For example, a 55-year old mother would be unique and her age would be suppressed. • To ensure primary suppressions cannot be derived by subtraction, additional cells may be selected for secondary suppression. • Cells in aggregate data, where the value of the cell is the same as a row/column total, should be suppressed if it is considered that it could lead to disclosure of an additional attribute.	• In order to protect any disclosive zeros, these will need to be suppressed. • Does not protect against disclosure by differencing. • Outliers, that could identify a person are removed. • Can be effective in hiding disclosive cells. • Need to be aware of actual intervals that are being implicitly published. • When releasing multiple tables, it may be possible to unpick suppressions even if this is not possible when considering each table on its own.	• Tables with suppressed cells are harder to extract information than the same tables with rounded values. • Depending on the extent of the suppression, it can introduce a high level of distortion in some types of analysis, as the suppression or loss of records is not completely at random. • Most information about suppressed cells will be lost. • Information loss will be high if more than a few suppressions are required. • Complex to implement optimally if more than a few suppressions are required, and particularly complex for linked tables. • Secondary suppressions will hide information in safe cells.	• Statistical disclosure control technique that can be implemented in various forms, whereby the data are only partially released. • Census outputs • Health data set is being disclosed to a researcher
Value suppression	• Suppression can be used for microdata where particular variables can be suppressed for particular cases. • E.g. In the case of a 16-year old widower with a child on the dataset the age might be suppressed. In effect it would be marked as missing data.	• Minimises risk of identification.	• Most information about suppressed cells will be lost. • Information loss will be high if more than a few suppressions are required. • Complex to implement optimally if more than a few suppressions are required, and particularly complex for linked table.	• Statistical reports

[23] The Anonymisation Decision-Making Framework

Technique	Details	Impact on risk	Impact on utility	Uses
Generalisation and grouping	• Data is grouped at a granularity that obscures unit records; including aggregation of data as well as more advanced techniques. • A release of data is said to have the k-anonymity property if the information for each person contained is common with at least k-1 individuals whose information also appears in the release. • Data is expressed in summary form by grouping related values into categories or ranges. • This can reduce disclosure risks by removing unit level identifiers and turning atypical records, which generally are most at risk, into typical records. • Essentially, grouping trades accuracy or ‘resolution’ for privacy, since any analysis on the grouped data cannot be more specific than what the grouping permits. • For grouping to work effectively, the groups must be defined by someone with relevant domain knowledge.	• Grouping can suffer from some of the same re-identification risks such as masking, when joining several data sets results in the possibility of re-identifying data that is not re-identifiable in isolation. Grouping does mitigate this problem to a certain extent by necessitating more joined data before re-identification becomes possible than plain masking does.	• Generalisation provides a lesser degree of granularity.	• Statistical data, e.g. for patient diagnosis with approximate age range and geographical location.

6.5 Use of controls and safeguards

Applying safeguards and controls to data can reduce the risk of re-identification and better preserve the utility or richness of the information being released, possibly more so than can be achieved for the same utility impact by modifying the data itself. It is difficult to determine whether data are de-identified or not without reference to the environment. If data is detailed and/or sensitive, controls would need to be put in place to limit access to authorised users working in a secure facility. If data has minimal detail and is not sensitive a less restricted access option could be considered.

Examples of controls and safeguards include:

- Including only the information necessary to achieve the intended purpose.
- Specifying who is permitted to access the information.
- Allowing access only within a controlled environment (which has been adequately secured e.g., through the Cyber Security Vulnerability Assessment) and stating how access is obtained.

  Please refer to Cyber Security and Information security classification framework (QGISCF)
- Ensuring that those given access to the de-identified information cannot access the original information.
- Make arrangements for the destruction or return of the information on completion of the project.
- Enabling an analysis of information rather than providing access to it, e.g. running an analysis of the information and providing the result rather than the raw information or stipulating what analyses may or may not be conducted and where it is to be carried out.
- Using an information sharing agreement or a memorandum of understanding to limit use and disclosure of information, including a prohibition on any attempt at re-identification and specifying that all analytical outputs must be approved by the agency before they are published.[24]
- Cost is an important factor to consider when applying security level controls to data.

6.6 Risk assessment

Before releasing de-identified information, it is important to assess whether the chosen de-identification techniques, and any safeguards and controls applied to the environment in which the information will be released, are appropriate to manage the risk of re-identification.

Note: The level of data treatment appropriate for authorised access in a controlled environment is unlikely to be sufficient for open and unrestricted public access.

Re-identification generally occurs through:

- Poor de-identification – where identifying information is inadvertently left in the information.
- Data linkage – it can be possible to re-identify individuals by linking de-identified information with an 'auxiliary dataset' that contains identifying information.

[24] Privacy and De-identification | Office of the Information Commissioner Queensland

- Pseudonym reversal – if an algorithm with a key is used to assign pseudonyms, it can be possible to use the key to reverse the pseudonymisation process to reveal identities.
- Inferential disclosure – this occurs when personal information can be inferred with a high degree of confidence from statistical attributes of the information.[25]

The following factors will need to be considered when determining if data is, or has been, sufficiently de-identified:

- the amount of information about the individual(s) available in other datasets and published information.
- the ease of access to source records/information.
- the level of detail provided and how current the information is.
- intimate knowledge (e.g. friends, family, minority groups, and small or tight-knit communities).
- the likelihood of re-identification, and
- any consequences that may result from the information being re-identified.[26]

An analysis is performed on relevant plausible scenarios for the data situation considering the how, who and why of a potential breach. A typical example is the nosy neighbour scenario where information could be relatively easily obtained by observation of an individual's neighbour. Someone might recognise information which could possibly pertain to a neighbour or fish for a neighbour's personal details in a dataset.[27] Penetration testing is used to validate assumptions, made during analysis of data situation scenarios, by simulating attacks using friendly and motivated antagonists.

6.6.1 Disclosure control processes

Disclosure control processes are implemented which consider either or both of the two elements of the data situation: the data and its environment. When the requirement for stronger controls is identified during a risk analysis there are two (2) (non-exclusive) choices:

- Reconfigure the data environment.
- Modify the data, including possibly reducing the amount of data under consideration.

After performing a risk assessment, a mitigation strategy should be considered for release issues:

- Mechanisms should be put in place to deal with disclosure, if it occurs.
- These should include a robust audit trail, a crisis management policy, and adequately trained staff.[28]

The following report outlines how agencies can manage privacy risks when releasing de-identified data Privacy and Public Data Audit Report.

[25] The De-Identification Decision-Making Framework: appendices (September 2017)
[26] Privacy and Right to Information Unit, Department of Health
[27] The De-Identification Decision-Making Framework: appendices (September 2017)
[28] CSIRO - The De-Identification Decision-Making Framework

7. Anonymise data

7.1 Definition:

Anonymisation is the permanent removal of identifying information, with no retention of the identifying information separately.[29]

7.2 Considerations

There are several techniques used to support the anonymisation process.

Table 4 lists the factors that should be considered for anonymisation.

TABLE 4 CONSIDERATIONS WHEN ANONYMISING DATA

Factor	Consideration
Population and sampling	• Who were the target population of the study and how was sampling conducted? • How many people belonging to the population were included in the sample? • What is known about the population beforehand (e.g. distribution of gender and age)? • Do individuals belonging to the population share a rare phenomenon?
Content of the data	• What kinds of direct and indirect identifiers do the data contain? • What combinations of information in the data could be used to identify an individual? • Does the dataset contain information related to third persons and can individuals be identified based on this information? • Does the dataset contain exceptional or unique information? • Does the dataset contain sensitive information?
Dataset age	• Have the data of the population in the dataset changed over time?
Information on the respondents available in other sources	• Is it possible to connect the information in the data to information from other sources? • Is it possible to identify individuals based on information available in other sources?
Usability vs. anonymity	• What types of information in the data are the most significant with regards to research, i.e. what information must be preserved during anonymisation and what information can be removed?

7.3 Process

The anonymisation process involves the following steps:

- Removing explicit identifying information about an individual (e.g. person's name, address, date of birth and unit record number).
- Applying expert statistical knowledge to render information not individually identifiable and to ensure that the risk is very small that the information could be used, alone or in combination with other information to identify an individual.[30]
- The use of controls and safeguards in the data access environment to prevent re-identification. Please see 6.5 Use of controls and safeguards for details.

[29] Australian National Data Service Guide: De-identification
[30] The Anonymisation Decision-Making Framework

7.4 Anonymisation techniques

Table 5 provides an overview of the more commonly used techniques (see Appendix E for additional anonymisation techniques) used to anonymise data for specific use cases, together with their associated impact on risk and data utility. Determination of which method is most appropriate for the data being released should be assessed by the expert on a case-by-case basis and will be guided by input from the health care providers.

TABLE 5 COMMONLY USED ANONYMISATION TECHNIQUES

Technique	Details	Impact on risk	Impact in utility	Uses
Remove direct identifiers	• Remove details that identify a person. • Appendix C: Direct identifiers	• Reduces risk of re-identification.	• Minimal data available to answer the question.	• Research statistics
Anonymisation techniques when deriving aggregate data				
Aggregation	• Used on indirect identifiers • An aggregation function is used to reduce many values down to a single value. • Typical aggregation functions include numeric calculation like a count, sum, maximum, minimum or average (typically a mean). • Calculations resulting in a TRUE or FALSE value can also be used in some programs. • Aggregate or reduce the precision of a variable, e.g. age or place of residence. • Provide 5-year age groups rather than date of birth. • Provide a metropolitan/rural indicator or statistical level area 2 (SA2) rather than a postcode and locality of residence. • Provide diagnosis related group instead of individual diagnosis and procedure codes. • No cell in the table(s) output may related to a single individual. • Ensure that the pool of people who could potentially have contributed to unit record data or to a cell in aggregate data is as large as possible while considering the data request. • Use of a numerical test, i.e. provision of unit record data for sub-groups where their estimated population is not less than a value set by the custodian.	• Attribute disclosure and disclosure by differencing are particular problems. • Aggregate data in any form can present an identification risk if individual responses can be estimated or derived from the output, e.g. outliers in a graph. • An identification risk exists if users have access to multiple tables that contain some common elements. • In magnitude tables, table cells present an identification (or disclosure) risk when they are dominated by values relating to one or two businesses or individuals.	• In certain cases there will be some loss of utility, in others analysis on the data may be significantly reduced.	• Cells to be published relate to population > 1,000 • Statistical tables • Admission counts • Count (frequency) tables – cells contain the number of individuals or organisations contributing to the cell (e.g. the number of people in various age groups) • Magnitude tables – cells contain values calculated from a numeric response (e.g. total income or profit). • Graphs • Maps • Aggregate data without small cell values.
Statistical disclosure control: the four related techniques described below may be used to transform small numbers, e.g. less than '5' appearing in cells within tables of aggregate data.				

Table redesign	• Small cell values (e.g. containing values between 1 and 4) are changed or removed.
	• No cell in the aggregate data output may be ≤ 5.
	• Disguise unsafe cells by:
	  • Grouping categories within a table
	  • Aggregating to a higher level geography or for a larger population sub-group
	  • Aggregating tables across a number of years/months/quarters
	  • Rounding
		• If unsafe cells remain in the output tabulation, further protection methods should be considered in order to disguise them, e.g. controlled rounding or cell suppression.	• Detail in the table will be reduced.
			• Original counts in the data are not damaged.
				• Cells to be published relate to population < 1,000 people
				• Statistical data
				• Census data
				• Population data

Cell modification – cell suppression	Refer to Cell suppression in: 6.4 De-identification techniques			
Cell modification – rounding	• Values in all cells in a table are adjusted to a specified base.	• Provides protection for zeroes. • Protects against disclosure by differencing and across linked tables.	• Counts are provided for all cells	• Statistical reports

Technique	Details	Impact on risk	Impact in utility	Uses
	• Small cells should be avoided by aggregating variables, e.g. age group ranges 65-74, 75-84, 85+ are replaced with 65+. • Data from small areas or communities should be aggregated over a number of years. • If this is not possible, then small cells may be suppressed. • Random rounding requires auditing; controlled rounding requires specialist software.	• Cannot be used to protect cells that are determined unsafe by a rule based on the number of statistical units contributing to a cell.	• Controlled rounding preserves the additivity of the table and can be applied to hierarchical data. • Uncertainty is created about the real value of any cell while slightly distorting the data.	
Cell modification – Barnardisation	• A post-tabular method for frequency tables where internal cells of every table are adjusted by +1, 0 or -1, according to probabilities.	• High level of adjustment may be required in order to disguise all unsafe cells. • Protects against disclosure by differencing.	• Will distort distributions in the data.	• Statistical data.
Anonymisation techniques when deriving individual-level data				
Data suppression	• Identify a sample of records, or certain data items in all records, and withhold these from the output. • Free-text data items, and human images, must be suppressed. • No other fixed standard test, but data reduction must be sufficient to make a significant contribution to anonymisation.	• Outliers, that could identify a person are removed.	• Detail in the table will be reduced.	• Statistical data.
Variable suppression	• Involves the removal or withholding of a data variable's values, e.g. removing name, address, postcode from an output. • All other variables in the record, i.e. those that are not quasi-identifiers, remain untouched.	• Reduces the risk of identification.	• It may not always be plausible to suppress some variables because that will reduce the utility of the data.	• Statistical data.
Record suppression	• If variable suppression and reduction in detail techniques do not adequately anonymise the data then the alternative is the removal and withholding of the data records that create a high re-ID risk.	• Reduces the risk of identification.	• Extensive suppression can introduce a high level of distortion in some types of analysis since the loss of records is not completely random and may reduce the usefulness.	• Statistical data.
Reduction in detail in indirect identifiers	• Identify, and withhold or transform indirect identifiers so they are less likely to reveal identity. • No date of birth, e.g. transform to age, year of birth, or 5-year age band. • No event dates, e.g. transform admission date to admission year, or month and year.	• Reduces the risk of identification.	• Output - individual-level data without indirect identifiers, or with indirect identifiers.	• Statistical data.
Suppression of direct identifiers	• Identify and withhold direct identifiers. • Suppression of name, address, widely-used unique person or record identifier (notably Medicare Number, Hospital Number), telephone number, email address, and any other data item that on its own could uniquely identify the individual.	• Removes or eliminates certain features about the data that could be identifying.	• Output – individual-level data without direct identifiers.	• Information on clinic sessions scheduled for a practice.
Metadata-level controls				
Sampling fraction	Refer to Sampling fraction in: 6.4 De-identification techniques			
Choice of variables	Refer to Choice of variables in: 6.4 De-identification techniques			
Level of detail	• Decisions over level of detail complement those over choice of variables. • Consider categories with small counts and determine whether merging them with other categories would significantly lower disclosure risk with minimal impact on data utility.	• Changing the detail on variables will reduce re-identification risk. • There is a reduction in key power. • If a variable has come categories that might be considered sensitive then sensitivity can be	• Do not distort the data and are transparent in their effects. • Impact on utility is similar but more subtle than the impact of removing whole variables.	• Statistical data.

Technique	Details	Impact on risk	Impact in utility	Uses
	• Variables such as geography and time are problematic. The area of residence is a highly visible component of an individual's identity, and so geographical detail should be constrained and data area released at a courser level. • Time-based variables, such as exact date of birth, can be identifying when combined with other variables and should be constrained.	reduced by merging these with other categories.	• Some variables can be more important than others. • Some aggregations will result in loss of utility.	
Perturbation or distorting the data				
Rounding	Refer to Rounding in: 6.4 De-identification techniques			

7.5 Risk assessment

Before releasing anonymised data, it is important to perform quality assurance on the chosen anonymisation techniques. An analysis is performed on relevant plausible scenarios for the data situation considering the how, who and why of a potential breach. A system of scenario analysis involving a classification scheme facilitates generation of a set of key variables that are likely to be available to an antagonist.

Inputs

- **Motivation** – What are the antagonists trying to achieve?
- **Means** – What resources (including other data) and skills do they have?
- **Opportunity** – How do they access the data?
- **Target variables** – For a disclosure to be meaningful something has to be learned, this is related to the notion of sensitivity.
- **Goals achievable by other means** – Is there a better way for the antagonists to get what they want than attacking the dataset.
- **Effect of data divergence (differences between datasets)** – All data contain errors/mismatches against reality. How will that affect the attack?

Intermediate outputs (to be used in the risk analysis)

- **Attack type** – What is the technical aspect of statistical/computational method used to attack the data?
- **Key variables** – What information from other data resources is going to be brought to bear in the attack?

Final outputs (the results of the risk analysis)

- **Likelihood of attempt** – Given the inputs, how likely is such an attack?
- **Likelihood of success** – If there is such an attack, how likely is it to succeed?
- **Consequences of attempt** – What happens next if they are successful (or Not)?
- **Effect of variations in the data situation** – By changing the data situation can you affect the above?

A typical example of a plausible scenario is the opportunistic targeting attack which considers an antagonist who is drawing on publicly available data sources, targeting a small number or individuals, who have visibility perhaps because of media coverage, without any resource constraints. Penetration testing is used to validate assumptions, made during analysis of data situation scenarios, by simulating attacks using friendly and motivated antagonists. [31]

In order to safeguard against disclosure please refer to 6.6.1 Disclosure control processes

[31] The Anonymisation Decision-Making Framework

8. Reference guidelines

- AIHW Guidelines for the Disclosure of Secondary Use Health Information for Statistical Reporting, Research and Analysis
- AIHW Data Governance Framework
- Australian National Data Service Guide: De-identification
- Australian National Data Service Guide - Publishing and sharing sensitive data
- CSIRO, The De-Identification Decision-Making Framework (September 2017)
- Definitions for identifiable, de-identified, non-identifiable, re-identified and anonymised data
- Department of Health, Privacy and Right to Information Unit 'Is it personal information' Fact Sheet
- Guidance on De-identification of Protected Health Information (November 2012)
- OAIC De-identification and the Privacy Act
- OIC Queensland
- The Anonymisation Decision-Making Framework
- The De-Identification Decision-Making Framework: appendices (September 2017)
- Tools for De-Identification of Personal Health Information: prepared for the Pan Canadian Health Information Privacy (HIP) Group (September 2009)

9. Legislation

- *Hospital and Health Boards Act 2011* (Qld) ss.139 - 142
- Information Privacy Act 2009 (Qld) NPP1-9
- *Private Health Facilities Act 1999* (Qld) s. 147,
- *Public Health Act 2005* (Qld) ss. 219 – 228, s. 230, ss. 237 – 249

Appendix A: Acronyms

Acronym	Description
ABS	Australian Bureau of Statistics
AIHW	Australian Institute of Health and Welfare
CSIRO	Commonwealth Scientific and Industrial Research Organisation
CT	Computed Tomography
DICOM	Digital Imaging and Communications in Medicine
HHB	*Hospital and Health Boards Act 2011*
HIS	Health Informatics Services
MRI	Magnetic Resonance Imaging
NMDS	National Minimum Data Set
OIC	Office of the Information Commissioner Queensland
PET	Positron Emission Tomography
PHRN	Population Health Research Network

Appendix B: Terms and Definitions

Terms	Definition	Source
Aggregate data	Aggregate data are produced by grouping information into categories and aggregating values within these categories. E.g. a count of the number of people of a particular age (obtained from the question 'In what year were you born?').	AIHW Guidelines for the Disclosure of Secondary Use Health Information for Statistical Reporting, Research and Analysis
Anonymised data	Data that have had identifying information permanently removed, with no retention of the identifying information kept separately.	Definitions for identifiable, de-identified, non-identifiable, re-identified and anonymised data
Antagonist	An antagonist is a person who might seek to re-identify an individual. They could be a malicious actor or simply a 'nosy neighbour'.	OIC, Privacy and Public Data Audit Report
Attribute disclosure (attribution)	This is the process of associating a particular piece of data with a particular population unit (person, household, business or other entity). In essence, it means that something new is learned about that population unit. Attribute disclosure often follows re-identification, however it can occur without re-identification.	CSIRO, The De-Identification Decision-Making Framework (September 2017)
Auxiliary information	Information, usually in the form of a dataset, that is available to the antagonist and is not contained within the target dataset. There are four sources of auxiliary information: • Datasets containing the same information for the same (or sufficiently similar population • Information that is publicly available (e.g. public registers or on social media) • Information obtained from local knowledge (e.g. house details obtained via an estate agent or by physical observation) • Information obtained through personal knowledge (e.g. things known about neighbours or work colleagues).	The Anonymisation Decision-Making Framework
Biometrics	These are codifications of unique, or statistically very likely to be unique, physical characteristics of individuals, to be used intentionally as identifiers, e.g. fingerprints, iris scans, gait	The De-Identification Decision-Making Framework: appendices (September 2017)

Terms	Definition	Source
	recognition systems, DNA and handwritten signatures.	
Confidential information	Means information, acquired by a person in the person's capacity as a designated person, from which a person who is receiving or has received a public sector health service could be identified.	*Hospital and Health Boards Act 2011* (Qld) s.139
Data	The representation of facts, concepts or instructions in a formalised (consistent and agreed) manner suitable for communication, interpretation or processing by human or automatic means. Typically comprised of numbers, words or images. The format and presentation of data may vary with the context in which it is used. Data is not information until it is utilised in a particular context for a particular purpose.	QGEA Glossary
Data key	A key which holds a variable value which can be applied to a string or a text block, in order for it to be encrypted or decrypted.	https://www.techopedia.com/definition/16080/data-key-cryptography
Data linkage	A process that compares records from one or more datasets with the objective of identifying pairs of records that correspond to the same population unit. Such pairs of records are said to be matched. This is also called statistical linkage, data linkage, or record linkage.	CSIRO, The De-Identification Decision-Making Framework (September 2017)
Data utility	A term describing the value of a given data release as an analytical resource – the key issue being whether the data represent whatever it is they are supposed to represent.	CSIRO, The De-Identification Decision-Making Framework (September 2017)
Dataset	Any collection of data about a defined set of entities, called population units, (whether persons, households, businesses, or other entities). Normally used to mean microdata (i.e. not summary/aggregate statistics).	CSIRO, The De-Identification Decision-Making Framework (September 2017)
De-identified data	De-identification involves removing or altering information that identifies an individual or is likely to enable their identification. De-identified data could potentially be re-identified.	Definitions for identifiable, de-identified, non-identifiable, re-identified and anonymised data
Direct identifier	Any data item that, on its own, could uniquely identify an individual case, such as a data subject's name,	The Anonymisation Decision-Making Framework

Terms	Definition	Source
	address and unique reference number, e.g. Medicare Card number.	
Disclosive zeros	Disclosure can arise from tables with larger values, where they appear in rows or columns dominated by zeros. Specific care should be taken if analysis shows that no one in a selected population has a particular attribute. This in itself can be disclosive about the selected population.	NHS National Services Scotland: Statistical Disclosure Control Protocol
Identifiable data	Data that can uniquely identify an individual. Examples of direct identifiers include name, address, driver's licence number, patient UR number and Medicare number.	Definitions for identifiable, de-identified, non-identifiable, re-identified and anonymised data
Indirect identifiers	Indirect identifiers (or quasi-identifiers) are typical of information that on its own is not enough to identify someone but, when linked with other available information, could be used to deduce the identity of a person.	Data Management Guidelines - Anonymisation and Personal Data | Data Archive
Inferential disclosure	An inferential disclosure occurs if the dissemination of a dataset enables the antagonist to obtain a better estimate for a confidential piece of information than would be possible without the data.	CSIRO, The De-Identification Decision-Making Framework (September 2017)
K-anonymisation	A dataset provides k-anonymity for the data subjects represented if the information for each person contained in the data set cannot be distinguished from at least k-1 individuals whose information also appears in the data set. E.g. a data set has 5-anonymity if, for every record in the data set that describe characteristics of a data subject, there are at least four other individuals also represented by records in the dataset who share the same characteristics described by the record.	The Anonymisation Decision-Making Framework
Key variables	A variable common to two (or more) datasets, which may therefore be used for linking records between them. Key variables are those for which auxiliary information on the data subjects is available to the data antagonist and which provide a hook into the target dataset, allowing individuals to be matched (see auxiliary information).	The Anonymisation Decision-Making Framework

Terms	Definition	Source
Microdata	A microdata set consists of a set of records containing information on individual data subjects. Each record may contain hundreds or even thousands of pieces of information.	CSIRO, The De-Identification Decision-Making Framework (September 2017)
Noise	Noise pertains to the stability of the data. Some data is very stable and possesses little variability, while other data swings wildly and unpredictably from one value to another. The degree of swing is the amount of noise.	Quora Data Science
Non-identifiable data	Non-identifiable data is data which have never been labelled with individual identifiers. A subset of non-identifiable data are those that can be linked with other data so it can be known they are about the same data subject, although the person's identity remains unknown.	Definitions for identifiable, de-identified, non-identifiable, re-identified and anonymised data
Non-zero cells	Having a positive or negative value; not equal to zero.	Lexico Dictionary
Outlier	An unusual value that is correctly reported but is not typical of the rest of the population.	Confidentiality - Glossary - Data.gov.au
Overimputation	Replacing real values in a micro-dataset with ones that have been generated through a statistical model.	The Anonymisation Decision-Making Framework
Personal information	Information or an opinion, including information or an opinion forming part of a database, whether true or not, and whether recorded in a material form or not, about an individual whose identity is apparent, or can reasonably be ascertained, from the information or opinion.	*Information Privacy Act 2009* (Qld)
Rainbow table	A rainbow table is a listing of all possible plaintext permutations of encrypted passwords specific to a given hash algorithm.	https://en.wikipedia.org/wiki/Rainbow_table
Re-identifiable data	Re-identifiable data, from which identifiers have been removed and replaced by a code, but it remains possible to re-identify a specific	Definitions for identifiable, de-identified, non-identifiable, re-identified and anonymised data

Terms	Definition	Source
	individual by, for example, using the code or linking different data sets.	
Salt	In cryptography, a salt is random data that is used as an additional input to a one-way function that hashes data, a password or passphrase.	Salt (cryptography) - Wikipedia
Sensitive information	Is a subset of personal information, and is given a higher level of protection under the *Information Privacy Act 2009* (Qld). Sensitive information includes a person's personal information about, but not limited to the following: race or ethnic origin, sexual preferences or practices, philosophical beliefs, membership of a professional or trade association, religious beliefs or associations, political opinions, membership or a political association, health information.	Department of Health, Privacy and Right to Information Unit 'Is it personal information' Fact Sheet
Standard	A document, established by consensus and approved by a recognised body, that provides, for common and repeated use, rules, guidelines or characteristics for activities or their results, aimed at the achievement of the optimum degree of order in a given context.	Common terminology for use in health policy and plans
Statistical linkage	Refers to a process that classifies pairs of records across different datasets as matched (deemed to correspond to the same population unit) or not matched (deemed not to correspond to the same population unit).	The Anonymisation Decision-Making Framework
Target variables	Object of interest to an antagonist, and thereby subject to attack. Applies to an individual, a record, a variable, some information or a dataset.	CSIRO, The De-Identification Decision-Making Framework (September 2017)
Unit record data	Refers to information relating to an individual person, such as name, sex, date of birth, date of cancer diagnosis and cancer type.	Cancer Institute NSW

Appendix C: Direct identifiers

The following list of identifiers has been compiled from the National Minimum Data Set (NMDS) and the Safe Harbor Standard.

The following demographic items pose a particular risk for an individual's identification:

- Names
- Alias or previous name
- Sex
- Country of birth
- Preferred language
- Indigenous status
- Marital status
- Age
- All elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older.
- Telephone numbers/fax numbers/email addresses
- Person identifiers/Tax File Number/Australian Business Number
- Centrelink numbers
- Medicare card or private insurance numbers
- Medical record numbers (e.g. Unit Record Number)
- Account numbers
- Certificate/licence numbers
- Vehicle identifiers and serial numbers, including license plate numbers
- Device identifiers and serial numbers
- Web Universal Resource Locators (URLs)
- Internet Protocol (IP) addresses
- Biometric identifiers, including finger and voice prints
- Full face photographs and any comparable images
- Any other unique identifying number, characteristic, or code
- Area of usual residence
- Establishment identifier (particularly for establishments with small catchment area)
- All geographic subdivisions smaller than a state, including street address, city, suburb, postcode
- Current address or last known address, and 2 previous addresses
- Current or last known employer.

The following data items, included in NMDSs covering particular types of health service events, pose a particular risk that they may enable further information to be disclosed about particular individuals who may be known or ascertained to be clients of a health service.

Admitted patient care NMDS

- Activity when injured
- Additional diagnosis
- Admission date
- Care type
- External cause – admitted patient
- Infant weight – neonate, stillborn
- Inter-hospital contracted patient
- Mode of separation (e.g. left against medical advice, died)
- Place of occurrence of external injury
- Principal diagnosis
- Procedure
- Separation date
- Source of referral to public psychiatric hospital (e.g. law enforcement agency)

Admitted patient mental health care NMDS

- Additional diagnosis
- Admission date
- Care type
- Mental health legal status
- Mode of separation (e.g. left against medical advice, died)
- Principal diagnosis
- Separation date
- Source of referral to public psychiatric hospital (e.g. law enforcement agency)

Admitted patient palliative care NMDS

- Additional diagnosis
- Admission date
- Care type
- Mode of separation (e.g. left against medical advice, died)
- Principal diagnosis
- Separation date

Alcohol and other drug treatment services NMDS

- Date of cessation of treatment for alcohol and other drugs
- Date of commencement of treatment for alcohol and other drugs
- Main treatment type for alcohol and other drugs
- Method of use for principal drug of concern
- Other drug of concern
- Other treatment type for alcohol and other drugs
- Reason of cessation of treatment for alcohol and other drugs

- Source of referral to alcohol and other drug treatment service

Community mental health care NMDS

- Mental health legal status
- Principal diagnosis
- Service contact date

Elective surgery waiting times NMDS

- Indicator procedure
- Listing date for care
- Reason for removal from elective surgery waiting list
- Date of removal

Injury surveillance NMDS

- Activity when injured
- Bodily location of main injury
- External cause –admitted patient
- External cause – human intent
- Narrative description of injury event (depending on the amount of detail provided)
- Nature of main injury
- Place of occurrence of external injury
- (Note: this list includes all items listed in the NMDS in Version 12 of the National Health Data Dictionary.)

Non-admitted patient emergency care NMDS

- Date patient presents
- Emergency department arrival mode
- Emergency department departure status
- Time patient presents
- (Note: this list is based on the items listed in the NMDS in Version 12 of the National Health Data Dictionary. Other dates and times, and diagnostic data items such as presenting problem, should be added as they are developed and endorsed for inclusion.)

Perinatal NMDS

- Actual place of birth (especially the non-hospital values of this data item)
- Birth order (especially for multiple births)
- Birth plurality (especially for multiple births)
- First day of last menstrual period
- Gestational age (especially low and high outliers)
- Infant weight – neonate, stillborn (especially low and high outliers)
- Method of birth
- Onset of labour
- Separation date
- Status of the baby (especially stillbirths).

Other

- Treating clinician
- Treating clinic

Appendix D: Additional de-identification techniques

Technique	Details	Impact on risk	Impact on utility	Uses
Data reduction (protects tables by either combining categories or suppressing cells)				
Data modification (changes all non-zero cells by a small amount without reducing the table's overall usefulness for most purposes)				
Data swapping	• Records of pairs of individuals of roughly the same characteristics of interest are identified. • The values of particular variables are then swapped between the two records. • As a result, a dataset is created with records that are no longer the original records but which, on aggregate analysis, will achieve the same results as would have been achieved using the original dataset. • E.g. a person from a particular town in Australia may speak a language that is unique in that town. Information about that individual's spoken language could be swapped with the spoken language information for another individual with otherwise similar characteristics (based on age, gender, income or other characteristics as appropriate) in an area where the language is more commonly spoken.	• Modification in the aggregate data will reduce the risk of subtraction attacks including foiling any attempt to link on the fine geography. • Increases uncertainty • It reduces risk where multiple data products are being released from a single data source, e.g. a sample of microdata with coarse geography (level 1) and aggregate population tables of counts for fine geography (level 2) is a common set of census outputs. • Modest data-swapping between level 2 areas within the level 1 areas means the microdata itself is unperturbed. • However, the perturbation in the aggregate data will reduce the risk of subtraction attacks and make any attempt to link the fine geography.	• Impact on data utility can be significant and it will often affect relationships between variables in an arbitrary and unpredictable manner. • Not used routinely in data situations where a single data product is involved. [32]	• Used in cases where a unique characteristic could identify a person, e.g. rare disease. • Produces overall aggregate statistics. • Used where multiple data products are being released from a single data source. • Used in census outputs.
Perturbation	• Numerical data can be protected by adding some randomly selected amount of noise (e.g. a random draw from a normal distribution with mean equal to zero). • Adding noise to values can reduce the possibilities of accurate matching on the perturbed data and distort the values of sensitive variables. • E.g. the values of age, country of birth, and sex might be switched for at-risk records with those for other records. • A patient's age may be reported as a random value within a 5-year window of the actual age.	• The data distortion is designed to increase the antagonist's uncertainty about any match, and so reduce the risk of re-identification. • Sensitive data potentially increases the risk and impact of disclosure.	• Maintains statistical properties about the original data, such as mean or variance.	• Randomisation of direct and indirect identifiers can generate realistic data for system testing without exposing person identifiers to vendors, implementers, system testers, and other third parties.33 • Statistical reports
Random rounding	• Small values are replaced with other small random numbers in a table. • Random rounding to base X involves randomly changing every number in a table to a multiple of X. • E.g. random rounding to base 3 (RR3) means that all values are rounded to the nearest multiple of 3. • Each value, including the totals, is rounded independently. • Values which are already a multiple of 3 are left unchanged.	• The original values cannot be known with certainty.	• Results in some data distortion so that the sum of cell values within or between tables will not equal the table total.	• Surveys • Count tables
Graduated random rounding	• Similar to random rounding. • After specialised cell sizes the rounding base increases. • A small number will have a smaller rounding base than a large number.	• Protection offered does not diminish for large-valued cells.	• Accuracy of reported data is slightly reduced.	• Magnitude tables (cells contain values calculated from a numeric response, e.g. total costings). • Count tables
Controlled rounding	• Is a procedure that perturbs tabular data. • It is constrained to have the sum of the cells equal to the appropriate row or column totals within a table.	• Reduces risk of statistical disclosure.	• May not provide consistency between tables.	• Statistical reports

[32] The Anonymisation Decision-Making Framework
[33] Tools for De-Identification of Personal Health Information: prepared for the Pan Canadian Health Information Privacy (HIP) Group (September 2009)

Technique	Details	Impact on risk	Impact on utility	Uses
K-anonymisation	• Works by guaranteeing that for a given set of key variables (X) there exists no combination of values (Xj) for which there are fewer that k data units; k is defined by the entity carrying out the anonymisation. • The general principle is that if a user knows fewer than k individuals with the attributes Xj then precise re-identification is prevented.[34] • An understanding of the data environment is required to determine the 'correct' level of k or the combinations of variables.	• The value for k should be set at a level that is appropriate to mitigate risk of identification by the anticipated recipient of the data set. • It does not protect against attribute disclosure. If a record shares key attributes with k-1 other data • units, that may not help if all k units share a value on some sensitive attribute. • Need to understand what k actually means for the data and how it relates to what the antagonist might be able to do. • L-diversity was introduced to resolve this problem.	• Generalisation provides a lesser degree of granularity.	• Statistical reports • Statistical data for patient diagnosis with approximate age range and geographical location.
K-anonymity – weak	• K = 3 • The variables, i.e. data items, not controlled through k-anonymity must exclude: - Any derivation of date of birth (such as age range) - Gender - Ethnic category - Postcode – a metropolitan/rural indicator or SA2 rather than postcode - Event dates (such as hospital admission date, whereas hospital admission month and year is acceptable). - Employer • Occupation or staff group	• Reduces the risk of identification. • It does not protect against attribute disclosure.	• Generalisation provides a lesser degree of granularity.	• Publication of low-level health related statistical data.
K-anonymity - strong	• K = 5 • All variables but one must be controlled through k- anonymity. • The uncontrolled variable should not be full date of birth or ethnic category. • There are two methods of reducing the granularity of a data set so that it satisfies the k-anonymity property: • Suppression: where sensitive values are removed or replaced with 'placeholder' symbols, e.g. replacing name and religion values with a star '*', and • Generalisation: where individual attribute values are replaced with a broader category, e.g. replacing precise ages with one of a fixed set of age ranges, age 25 becoming 'between 20 and 30'.	• Reduces the risk of identification. • It does not protect against attribute disclosure.	• Generalisation provides a lesser degree of granularity.	• Publication of high-level data, e.g. HIV statistics
L-diversity	• Deals with attribute disclosure in k-anonymity by imposing a further constraint. • Each equivalence class (group of data units sharing the same attributes) must have multiple values on any variable that is defined as sensitive (target variable). • There has to be at least l different values for each sensitive variable within each equivalence class on the key variables.	• Need to understand what (l) actually means for the data and how this relates to what the antagonist might be able to do. • This technique is susceptible to many kinds of attacks especially when background knowledge is available to the attacker. • Risk that arbitrary decisions are made using the privacy model rather than the data situation.	• Can lead to counterintuitive outcomes.	• Statistical data
Differential privacy	• Current state of the art standard for controlling re-identification risk, that greatly reduces the disclosure risk issues in k-anonymity, L-diversity and other extensions. • An algorithm is said to be differentially private if the operation of that algorithm on the data set produces 'essentially the same' answers regardless of the presence or absence of any particular unit record.	• Limits the risk of re-identification.	• The main drawback of using differentially private algorithms for data analysis is that, like the other perturbation methods, it does not preserve the accuracy of the algorithms applied to the original unit records, and in fact will	• Description of patterns of groups within a dataset • Statistical data

[34] The Anonymisation Decision-Making Framework

Technique	Details	Impact on risk	Impact on utility	Uses
			deviate more from the results based on the original unit record data in proportion to the very privacy guarantees that it provides.	
Synthetic data generation	• Synthetic data is data that is created based on user-specified parameters to resemble the properties of data from real-world scenarios. • It is 'starting from scratch' to create new data based off a generalised statistical approach. • Synthetic data can be generated in a number of ways: - Use a generative computer model or generated adversarial network (GAN) to create a set of data points that cannot be differentiated from the real data. - In generative models, algorithms are fed with smaller real-world data which then gets derived by the algorithms to create similar data. - Enhanced sampling: either over-sample the minority class or under-sample the majority case to create a synthetic distribution of data. - Agent based simulation: use a simulation process where agents are developed to represent real-world entities that interact with each other, and these interactions are observed and measured to generate data. • There are two approaches to creating synthetic data: - **Partially synthetic data:** only data that is sensitive is replaced with synthetic data. - **Fully synthetic data** involves replacing an entire data set (rather than just a subset) with synthesised replacement data.	• In a test environment, synthetic data allows systems to be tested with data in a realistic way with less risk of re-identification. • It is superior to anonymisation and pseudonymisation of real data, which can be vulnerable to re-identification through cross referencing data sets. • With partially synthetic data some disclosure is possible owing to the true values that remain within the dataset.	• Used in place of real data when real data is incomplete or unavailable due to privacy restrictions. • Reduces the need to use production patient data for research purposes, alleviating privacy concerns. • It is superior to k-anonymisation in that it does not compromise the granularity of the original data through aggregation or removal. • Synthetic data is dependent on the model created to derive it. If the initial model is not sufficient to create quality synthetic data, the data created will be less reflective of authentic data.	• Use in test programs to detect fraud • Increasing efficiency and comprehensiveness of testing • Testing data and ensuring data quality • Conducting research without the use of personal, sensitive or confidential information. • Creation and enhancement of training data for machine learning and artificial intelligence tools through the provision of vast amounts of data. • Assists in providing data for training, learning and development initiatives.
Encryption or 'hashing' of identifiers	• Uses techniques that will obscure the original identifier, rather than remove it altogether. • Involves applying a hashing algorithm to a data item to scramble the identifier in a deterministic manner. • The pre-fix or salt (key word or phrase) is used to catenate with the data before it is hashed. • A Rainbow table is a precomputed table which stores pre-hashed data – this could potentially be used for malicious re-identification.	• Reduces the risk of identification	• Records can be joined together as long as the same salt is used.	• Used to link different datasets together (but without sharing the information in an identified form). • Request to know that multiple episodes relate to the same person in the same hospital, then the unit record numbers provided should be encrypted. • Productive test environment
Masking personal identifiers	• Remove or replace fields that may identify individuals, such as names, addresses, telephone numbers. • Sensitive information is replaced with realistic but inauthentic data • May involve suppressing entire fields or just at-risk data values. • Masking is often interpreted to mean randomisation, which involves replacing actual values with random values selected from a large database. • Use of database of first and last names, e.g. to randomise those fields. • Also generate random social security numbers to replace original ones.	• Masked data can still be identifying, particularly in combination with other data sets. • Data that may not be personally identifying in the context of a particular database can become so when joined with data in another data set. Assess re-identification risks and implement additional controls when other data sets are joined.	• The structure and functional usability of the data is retained while information that could lead to identification of an individual, either directly or indirectly, is concealed.	• Used to protect individual privacy in public reports • Can serve as a useful alternative when real data are not required, such as user training or software demonstration.
Pseudonymisation	• Mask personal information by replacing it with a pseudonym, a specially crafted value that can be used to identify unit records but does not itself contain personal information. Unique, artificial pseudonyms replace direct identifiers, e.g. Jo Bloggs = 5L7T LX619Z [unique sequence not used anywhere else]. • Pseudonymous IDs are provided by a custodian of the data. • Pseudo IDs can be provided to end-users that enable linking of information (e.g. from datasets collected over time) while still protecting the identity of those individuals.	• Use of the same pseudonym across multiple data sets and the availability of quasi-identifiers could leave pseudonymised data at risk of re-identification through linkage. • Pseudonymisation can be unpicked by creating a look-up between the un-pseudonymised and pseudonymised values. This could occur through clear and	• A pseudonym links de-identified data to the same person across multiple data records or information system without revealing the identity of the person and is therefore good for tracking.	• Where more granular information is required for richer analysis • Secondary use of clinical data • Clinical trials • Post marketing surveillance • Confidential patient-safety reporting (e.g. adverse drug effects)

Technique	Details	Impact on risk	Impact on utility	Uses
	• Pseudonymisation can: - Map a given direct identifier to the same pseudonymous ID - Map a given direct identifier to different pseudonymous IDs in a way that is context dependent (e.g. by assigning different pseudo IDs to different researchers or research institutions) - Map a given direct identifier to different pseudonymous IDs in a way that is location dependent (e.g. by assigning different pseudo IDS to data that comes from different data sources. • Pseudonymisation techniques should be carefully considered and implemented, as commonly used techniques only pseudonymise directly identifying information, while leaving quasi-identifiers in raw form. • Pseudonymisation can be performed with or without a possibility of re-identifying the data subject, i.e. use of reversible or irreversible pseudonymisation.	pseudonymised fields coming into contact with each other or through a pseudonymisation key or salt becoming available and used to generate a look-up between clear and pseudonymised data. • There should be a secure key management function.		• Comparative quality indicator reporting • Peer review • Equipment maintenance • Health research where a consistently applied pseudonymous identifier (it need not be reversible) allows for the tracking of patients over an extended period of time. • Health system planning • Public health surveillance – re-identification may be required for contacting data subjects, e.g. for management of disease outbreaks. • System testing

Appendix E: Additional anonymisation techniques

Technique	Details	Impact on risk	Impact in utility	Uses
Anonymisation techniques when deriving individual-level data				
Encryption	• Unit record numbers are encrypted to protect person identity	• Reduces risk of re-identification	• Encryption enables research requirement to be answered without the risk of identification.	Requirement of data request to know that multiple episodes relate to the same person in the same hospital
K-anonymisation	Refer to K-anonymisaton in: Appendix D: Additional de-identification techniques			
Perturbation or distorting the data				
Data swapping	Refer to Data swapping in: Appendix D: Additional de-identification techniques			
Overimputation	• Real values are replaced with ones that have been generated through a model. • In order for this to work without badly distorting the data, it may be necessary to allow the original values to be modelled back in. • A critical decision when overimputing will be what the user is told. • There is an option to tell the user any of the following: - that the data has been overimputed - how many values have been imputed - the model that has been used to do that imputation - the actual values that have been imputed. • Overimputation is also a good option if imputation is already being used to deal with missing values.	• The level of risk is dependent on the mechanism that is used to decide on the new value. • How transparent a custodian of the data is in divulging information about the overimputation • How much overimputation has been conducted.	• Is dependent on how good a model has been used to produce the overimputed values.	• Statistical reports