ACT Government | ACT Health

POLICY DOCUMENT

Statistical Disclosure Control Policy and Procedure

Document number	**AHDPD-04:2023**
Effective date	10 August 2023
Review date	10 August 2024
Author branch	Data Analytics Branch
Endorsed by	Executive Board (Operational)
Audience	All staff and contractors
Version number	1.0

Contents

Policy Statement

High quality data can provide valuable insights to assist the efforts of ACT Health Directorate (ACTHD) to improve health outcomes for the ACT community. Data can, for example, be used to quantify health system performance that, in turn, informs service delivery improvements or whole-of-Territory strategies and activities; or, to underpin clinical or epidemiological research.

There are obligations for government organisations, including ACTHD, to disclose (make data available) to authorised internal and external users, where public benefits can be realised.[1, 2] Organisations have an ethical and legal responsibility to ensure that the data they share are managed in a way that protects the privacy and confidentiality of data providers (an individual, household, business or other entity that supplies data[3]) captured within a data release.

Context

Legislation prescribes how organisations must manage and protect the confidentiality and security of the data they collect. This legislation includes the *Health Records (Privacy and Access) Act 1997* (ACT) and the *Information Privacy Act 2014* (ACT). Furthermore, the Australian community expects that any information provided to the government about them will be secure and safe from unauthorised access or disclosure.[4]

In the context of a Five Safes (Data Sharing Principles) disclosure risk assessment framework (the best practice approach to assessing and managing risks associated with data sharing and release), and as outlined in the *Data Disclosure Policy,* the application of statistical disclosure controls is a critical safeguard used to protect data provider confidentiality (the 'Data Principle').[5]

When data custodians approve the disclosure of **unit record data** (or **microdata**: each record contains information about a data provider), or **aggregated data** (information grouped into categories where values are combined) [6] to internal or external recipients, they must comply with relevant legislation, policies and agreements. They must ensure that data provider confidentiality is protected, and that the disclosure is lawful.

A data breach occurs where personal information is accessed or disclosed without authority, or is lost.[7] In the event of a data breach, aside from the timely management of breaches according to policies, organisations must take steps to ensure the security and privacy of individuals included in the release. This requires embedding procedures to 'render [the] data files more harmless'[8,9] through appropriate data management prior to disclosure.

[1] Proactive release of data (Open Data) Policy
[2] Data Availability and Transparency Act 2022
[3] Australian Bureau of Statistics Glossary
[4] OAIC Australian Community Attitudes to Privacy Survey 2020
[5] Department of the Prime Minister and Cabinet. *Best practice guide to applying data sharing principles.* 2019
[6] Australian Bureau of Statistics Glossary
[7] Office of the Australian Information Commissioner. *Part 1: Data breaches and the Australian Privacy Act*
[8] GDPR.EU How does the GDPR affect email? 2022
[9] ProtonMail. *Everything you need to know about GDPR compliance and email security*. 2018

Purpose

This *Statistical Disclosure Control Policy and Procedure* describes the policy rationale for the management of aggregate or unit record data that is authorised to be disclosed to internal or external recipients and the:

1. techniques that should be implemented by data users to limit the risk of an individual, organisation or other entity being directly or indirectly identified. This policy should be applied where aggregate data or unit record data are disclosed to internal or external parties through approved data sharing or release activities.

2. safeguards that are to be implemented prior to the transfer of information to help minimise harm to individuals or organisations should personal information, personal health information or other confidential information be disclosed to unauthorised recipients.

This policy should be read in conjunction with the *Transferring Confidential Data Policy and Procedure* that prescribes 1) the necessary authorisations and approvals; and 2) the modes of transfer that are to be used to securely transmit confidential information (including aggregate or unit record data) to internal or external recipients.

The *Data Breach Policy and Procedure* outlines the steps to be undertaken should a data breach be identified.

The *Data Disclosure Policy and Procedure* provides guidance to ensure appropriate approvals and other protections are implemented prior to disclosing data.

Scope

This policy applies to all ACTHD workers, including permanent, temporary, and casual employees, external contractors, consultants, students and volunteers. ACTHD workers are required to comply with all obligations in relation to privacy and data protection as directed in relevant legislation, agreements and policies.

This policy covers all data held by the ACTHD. It includes data or information acquired from external sources or provided to ACTHD by external data custodians or other data providers that are accessible through ACTHD systems.

This policy focuses on the management of data to prevent the re-identification of individuals or organisations included within a data disclosure. This policy does not address the conditions that must be met prior to any data sharing or release, such as data custodian or ethics committee approvals. This is addressed in the *Data Disclosure Policy and Procedure.*

Background

Under certain conditions, data held by ACTHD can be disclosed, through sharing or release, to authorised recipients. Data are:

- **shared** where they are made 'available to another agency, organisation or person under agreed conditions.'
- **released** where they are 'made publicly available with no or few restrictions on who may access the data and what they may do with it'.
  - An example is the ACT Open Data Portal. In this setting, data custodians cannot control who accesses the data or the purposes for which the data are used.[10]

Confidentiality and disclosure

In the ACT, personal information (including sensitive information) is that defined under the *Information Privacy Act 2014 (ACT)*. Personal health information is as defined under the *Health Records (Privacy and Access) Act 1997 (ACT)* – see Glossary. Examples of personal information and personal health information can be found within the legislation, and in the *Transferring Confidential Information Policy and Procedure* or the *Data Breach Policy and Procedure*.

The Australian Bureau of Statistics defines **confidentiality** as 'protecting the secrecy and privacy of information collected from individuals and organisations and ensuring that no data are released in a manner likely to enable their identification'.[11]

'**Disclosure**' (or '**re-identification'** or a '**breach of confidentiality'**) occurs when previously unknown information about a person is revealed in a data release.[12] When sharing data, certain safeguards or **disclosure controls** (the process of limiting the risk of an individual or organisation being directly or indirectly identified'[13]) may need to be implemented to ensure confidentiality.

When sharing data, certain safeguards or disclosure controls may need to be implemented to ensure confidentiality.

Disclosed data may contain:

- **direct identifiers** - information that unambiguously identifies an individual, or an entity such as a business or health care service.[14] Direct identifiers are personal information as defined under the *Information Privacy Act 2014* (ACT). Examples include organisation or individual names and addresses.

[10] Department of the Prime Minister and Cabinet. *Best practice guide to applying data sharing principles*. 2019

[11] Australian Bureau of Statistics Glossary

[12] International Household Survey Network, 2014. *Introduction to Statistical Disclosure Control.* ISHN Working Paper No. 007

[13] Australian Bureau of Statistics Glossary

[14] International Household Survey Network, 2014. *Introduction to Statistical Disclosure Control.* ISHN Working Paper No. 007

- **quasi-identifiers** (or **implicit identifiers**) - other information, aside from direct identifiers, that can be connected to additional information (for example the internet or data held in a private collection) to re-identify individuals, organisations (including businesses) in a 'de-identified' record.[15] As such, where direct identifiers have been removed from a record, it may be possible to re-identify a person where quasi-identifiers remain.

There are two main types of identification risks[16] that can occur when data are made available to others:

1. **Identity disclosure**:
    a. **Direct identification**: Where a data release contains direct identifier/s that establishes the identity of a person, group or organisation.
    b. **Indirect identification**: Where 'the identity of a person, group or organisation is disclosed due to a unique combination of characteristics (that are not direct identifiers) in a dataset'.
        - An example is a celebrity identified from data containing age, sex, occupation and income.

    Identification can be:
    - **Spontaneous**: This is a non-deliberate identification. It can occur where there are individuals with a rare characteristic within the data. An example would be a 99-year-old man who lives in a rural postcode.
    - **Deliberate:** Where the data recipient crossmatches or links unique characteristics that are common to the released data and other information such as that found on the internet or already held; or tries to identify a record that contains specific characteristics known to that person.[17]

2. **Attribute disclosure** occurs where 'previously unknown information is revealed about an individual, group or organisation (without necessarily formally re-identifying them)'.[18]
    - For example, if an income bracket for all women aged 60 to 65 years living in a small geographic area were reported, then the income for any woman of that age living in that area would be known to an outsider.

Section 18 of the *Information Privacy Act 2014* (ACT) states that personal information is **de-identified** if 'the information is no longer about an identifiable individual or an individual who is reasonably identifiable'. De-identification requires the:

1. 'removal or alteration of other information that could potentially be used to re-identify an individual, and/or the
2. use of controls and safeguards in the data access environment to prevent re-identification'.[19]

[15] International Household Survey Network. *Introduction to statistical disclosure control*. IHSN Working Paper No. 007, 2014
[16] Australian Bureau of Statistics Glossary
[17] Australian Government Data.gov.au *Part 5 - Managing the risk of disclosure: Treating Microdata*
[18] Australian Bureau of Statistics Glossary
[19] Office of the Australian Information Commissioner 'De-identification and the Privacy Act'

In a de-identified record, there are several statistical controls[20] that can be applied to help minimise the risk of accidental or purposeful re-identification. Application of these controls can facilitate the release of data that would otherwise be inaccessible to authorised users, or release without the use of a secure access environment.

These controls can be applied to unit record data or to cells within aggregated data tables where counts or frequencies are small. Greater controls may need to be applied where the data contains sensitive personal information such as sexual orientation or ethnicity, or sensitive health information (for example mental health conditions, or communicable diseases). Techniques to manage aggregate and unit record data prior to release to help mitigate disclosure risks are outlined below.

In a de-identified record, there are several statistical controls that can be applied to help minimise the risk of accidental or purposeful re-identification.

Aggregate data disclosure risk control

Disclosure risk assessment

Prior to any sharing or release of aggregate data, a disclosure risk assessment should be undertaken. A case-by-case approach, considering any data custodian requirements, is recommended.

The aim of implementing statistical disclosure controls to a data release is to balance disclosure risk while retaining the usefulness or utility of the data for the data user. Confidentiality rules for aggregate data that may be determined by the data custodian:

1. **Frequency rule**: Application of a threshold value for the minimum number of contributions within a cell.
2. **Cell dominance rule**: Where 'a small number of data providers contribute to the cell total'.[21] This rule is designed to prevent the re-identification of individuals or organisations that contribute a large percentage of a cell's total value. An example would be where income or turnover is reported.

Control mechanisms include **data reduction methods,** such as collapsing categories and suppression. **Data modification** involves making changes to the data. The Australian Bureau of Statistics recommends beginning with the simpler data reduction techniques and implementing data modification techniques where disclosure risks remain.

Refer to the *Data Disclosure Policy* for detailed information about the approvals and assessments that need to occur prior to release or sharing.

[20] Australian Government Data.gov.au *Part 5 - Managing the risk of disclosure: Treating Microdata*
[21] Australian Government Data.gov.au *Part 4 - Managing the risk of disclosure: Treating Aggregate Data.*

Data reduction methods

To maintain confidentiality in a data table, combining or collapsing categories or suppression techniques may be used.

Combining or collapsing categories

This disclosure control method involves combining several response categories into one or reducing the amount of classificatory detail available in a table. Examples include grouping postcodes into greater geographical areas (such as Statistical Area Level 2 [SA2]), combining income ranges, or combining ages into broader groups so there are larger frequency counts within cells and outliers are hidden.

Suppression

Suppression involves withholding small counts considered to be a disclosure risk within a cell (or more than one cell where row or column totals are reported in a table). Suppression can be applied to aggregated tables, figures, maps, dashboards and data cubes. Consider whether the suppressed information has already been released elsewhere so that it remains possible to calculate the suppressed value from that information.

Suppression of data with small numbers applies where there:

- are small cell counts
- is a risk an individual could be identified, possibly leading to revealing information about them that was previously unknown
- is a risk of exposing an organisation's commercial operations.

In some circumstances small cell numbers can be released. These include:

- where there is approval from the data custodian
- where the data are a count of organisations such as the number of hospitals in the ACT
- for 'unknown' categories
- for operational data in dashboards for internal use only
- for specific reports such as perinatal deaths (with approval).

Consequential (or **secondary suppression**) may also be required where cells are suppressed and it is possible to calculate the withheld value from other values within the table, such as row or column totals. Consequential suppression should be carefully applied to permit the desired level of ambiguity for the small cells with the least amount of suppression.

Suppression procedure

Where there is a disclosure risk, small cell data should be replaced with an 'n.p. ('not provided' or 'not publishable')' or '<5'. This is referred to as **primary suppression**.

Unless there is specific agreement with data custodians, at ACTHD cells must not contain counts greater than 0 but less than 5.

Unless there is specific agreement with data custodians, cells must not contain counts greater than 0 but less than 5.

Suppression of cells with a zero

Suppression of cells that contain a zero is not usually required, on the basis that a count of no events is unlikely to be a threat to confidentiality. Sometimes, however, zero cells (cells with no contributors or all values were zero) or cells where 100% of respondents shared the same characteristics can pose confidentiality problems such as attribute disclosure. For example, Youth Survey data for a specific school containing a count of '0' for 'drugs never used' would indicate that all students used drugs and would lead to attribute disclosure for students known to attend that school. The following suppression examples are taken from the Australian Bureau of Statistics.[22]

Step 1: Identify the small numbers in the data to be released.

Example: Table 1 includes small numbers (a '3' and a '4').

Table 1: Number of people by age group (years) and income status

Age group	Low income	Medium income	High income	Total
15-19	16	0	0	16
20-24	8	10	7	25
25-29	**3**	8	11	22
30-34	**4**	5	18	27
Total	**31**	**23**	**36**	**90**

Step 2: Consider identification or attribute disclosure risks should no cells be suppressed.

Step 3: Seek approval to release small numbers, if appropriate.

Step 4: Suppress small numbers.

Example: In Table 2, the '3' and '4' have been suppressed, and have been replaced with an 'n.p.'

Table 2: Number of people by age group (years) and income status

Age group	Low income	Medium income	High income	Total
15-19	16	0	0	16
20-24	8	10	7	25
25-29	**n.p.**	8	11	22
30-34	**n.p.**	5	18	27
Total	**31**	**23**	**36**	**90**

[22] Australian Bureau of Statistics. *Treating aggregate data*

Step 5: Assess residual risk. Table 2 requires further suppression, as the value of the suppressed cells can be calculated using the row totals. For example, for the 30-34 age group, 27-5-18 = 4 (the value of the suppressed cell). Apply **consequential** (or secondary) **suppression** – the suppression of additional cells (Table 3).

Table 3: Number of people by age group (years) and income status

Age group	Low income	Medium income	High income	Total
15-19	16	0	0	16
20-24	8	10	7	25
25-29	**n.p.**	**n.p.**	11	22
30-34	**n.p.**	**n.p.**	18	27
Total	**31**	**23**	**36**	**90**

Data modification

Data modification involves changing the data in some way, or the purposeful introduction of random error into the data to maintain confidentiality. It is generally applied to non-zero cells in a table.

- **Rounding:** Examples include rounding a value up or down to a specified base (all values divisible by the same number such as 3, 5 or 10).

- **Perturbation**: There are changes (often random) to some or all non-zero cells with small numbers in a table. For frequency tables, a randomised number may be added to the original values. For magnitude data, the original values are multiplied by a random number.

Where data have been modified, an explanatory footnote should accompany the disclosure, so the recipient is aware that alterations have occurred.

Unit record data disclosure risk control

Disclosure risk assessment

A disclosure risk assessment should be performed prior to unit record data sharing.

The extent of statistical controls required to minimise disclosure risk for unit record data release depends on an assessment of the environment in which the data will be accessed. For example, where the data are released into a secure controlled access environment, such as the Australian Bureau of Statistics' DataLab, there is a trusted vetting of other information that can be viewed in tandem with the released data and an evaluation of outputs such as frequency tables or regressions. Use of this type of access environment supports the release of data that has been treated with fewer controls so that there is greater utility for the data user.

Agencies hosting secure access environments, such as the Australian Institute of Health and Welfare or the Australian Bureau of Statistics provide instructions to data users and custodians regarding the uploading of data to these settings.

Where the access or analysis environment cannot be controlled (i.e. the data is released directly to a researcher for access on a university server), the management of the data prior to release may need to be more robust. Prior to unit record data release, a disclosure risk assessment should consider:

- the possible intentions of the data recipient or user.
  - For example, has a human research ethics committee considered the project? Is a data sharing agreement in place? What are the credentials and affiliations of the data requester?
- whether rare characteristics (such as rare diseases, outliers by age) are included
- the degree of detail – increased detail in variables or the number of fields can increase identification risk
- the time period – more recent data may be associated with greater disclosure risk
- completeness – could a sample of the data serve the data user's purposes instead of the entire population?[23]

Refer to the *Data Disclosure Policy* for detailed information about the approvals and assessments that need to occur prior to release or sharing.

Unit record data disclosure risk management

Disclosure risk management for unit record data may involve the application of data perturbation (such as introducing random error or suppression) or data reduction methods (such as combining categories such as age, income, health service type or small geographic areas). Other methods are outlined below. ACTHD Epidemiology teams can provide advice on the best option if unit record data disclosure risk management is required.

If you need guidance or are unsure, seek expert assistance from Epidemiology teams.

Anonymisation

Direct and indirect personal identifiers are removed from the record so that the individual or organisation is no longer identifiable or re-identifiable.[24]

Pseudonymisation

Pseudonymisation is the 'processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately'.[25]

Pseudonymisation involves replacing identifying fields with one or more pseudonyms, or fictional identifiers. It is reversible and allows for re-identification later, should it be required. For example, a unique hospital patient identifier is replaced with another randomly generated identifier. A concordance 'map' is retained by the data custodian so that the original identifier remains known.

[23] Australian Government Data.gov.au *Part 5 - Managing the risk of disclosure: Treating Microdata*
[24] University College London. *Anonymisation and pseudonymisation*. 2022
[25] GDPR.EU Art. 4 GDPR Definitions

Data minimisation

Consider whether all the information that is intended to be sent is necessary to meet the data requester's needs. Limit the information to the bare minimum required for the intended purpose. The provision of month and year, instead of a full event date should be considered to minimise re-identification risk. Similarly, creating a variable of 'days from first event' rather than providing an event date may be sufficient to meet the data requester's purposes.[26]

Free text fields

Free text (string) fields can contain confidential information. Unless authorised by the data custodian, or in instances where the file is so small that the content can be easily checked for disclosure risks, do not include any free text fields in a released data file.

Unless authorised by the data custodian, or in instances where the file is so small that the content can be easily checked, do not include any free text fields in a released data file.

Roles and Responsibilities

Position	Responsibility
All staff	Adherence to this and related policies and procedures.
Data custodians	Provide guidance regarding the level of controls required to manage re-identification risks for data releases. Accountable for data governance decisions for assigned data or data sets and for authorising and facilitating safe data access, use and sharing.
Managers and executives	Ensure sufficient support and resourcing to implement robust data release assessment and data management.

Evaluation

Outcome Measures	Method	Responsibility
Where there is authorised or unauthorised access or disclosure of data, sufficient disclosure risk controls are applied to minimise re-identification risks.	Data Breach Register reports	Director, Data Strategy and Governance

[26] National Health Information Standards and Statistics Committee (NHISSC) 2017. *Guidelines for the Disclosure of Secondary Use Health Information for Statistical Reporting, Research and Analysis*

Related Documents

Legislation

- *Health Records (Privacy and Access) Act 1997* (ACT)
- *Information Privacy Act 2014* (ACT)
- *Privacy Act 1988* (Cwth)

Supporting Documents

- *Data Breach Policy and Procedure* [AHDPD-81:2021]
- *ACT Health Directorate Information Privacy Policy*
- CSIRO Data 61, 2017 *The de-identification decision-making framework*

Glossary

Term	Definition
Consumer – as defined in the *Health Records (Privacy and Access) Act 1997 (ACT)*	An individual who uses, or has used, a health service.
Personal health information – as defined in the *Health Records (Privacy and Access) Act 1997 (ACT)*	Personal health information, of a consumer, means any personal information, whether or not recorded in a health record— (a) relating to the health, an illness or a disability of the consumer; or (b) collected by a health service provider in relation to the health, an illness or a disability of the consumer.
Personal information – as defined in the *Information Privacy Act 2014 (ACT)*	Information or an opinion about an identified individual, or an individual who is reasonably identifiable— (i) whether the information or opinion is true or not; and (ii) whether the information or opinion is recorded in a material form or not; but (b) does not include personal health information about the individual.
Sensitive information – as defined in the *Information Privacy Act 2014 (ACT)*	Sensitive information, in relation to an individual, means personal information that is— (a) about the individual's— (i) racial or ethnic origin; or (ii) political opinions; or (iii) membership of a political association; or (iv) religious beliefs or affiliations; or (v) philosophical beliefs; or (vi) membership of a professional or trade association; or (vii) membership of a trade union; or (viii) sexual orientation or practices; or (ix) criminal record; or (b) genetic information about the individual; or (c) biometric information about the individual that is to be used for the purpose of automated biometric verification or biometric identification; or (d) a biometric template that relates to the individual.

Version Control

Version	Date	Comments
V0.1	October 2022	Initial draft. Data Strategy and Governance
V0.2	November 2022	Population Health, DAB
V0.3	March 2023	Legal Policy review
V 1.0	June 2023	Consultation feedback