Guidelines for the Use and Disclosure of Health Data for Statistical Purposes

Statistical Information Management Committee

March 2007

Table of contents

Introduction

These guidelines were developed by the Statistical Information Management Committee (SIMC) in response to concerns expressed by SIMC members about the use and disclosure of data. The guidelines were developed by a working group established by the SIMC in March 2005.

The purpose of these guidelines is to document what the SIMC considers to be good practice that all parties to the National Health Information Agreement (NHIA) can use to guide decision-making about the disclosure of health statistical data. They are intended to provide general guidance to assist data custodians to manage some general risks regarding the identification of individual patients and health service providers . However, they do not represent an attempt to isolate every possible risk which may occur in relation to every request. There will still be some variation in practice between jurisdictions in providing information, for example with respect to providing information about individual health service providers such as public hospitals.

These health-specific guidelines are based on the general techniques for confidentialising statistical information produced from administrative sources described in Appendix 4 of the National Statistical Service Handbook, which is available on the National Statistical Service website at:

http://www.nss.gov.au/nss/home.NSF/pages/NSS+Handbook?OpenDocument#Appendix%204

Appendix 4 of the Handbook also provides a description of the legislative privacy framework established by the Commonwealth *Privacy Act 1988*, together with links to relevant Commonwealth, State and Territory legislation and guidelines including the *Guidelines under Section 95 of the Privacy Act 1988* issued by the National Health and Medical Research Council (NHMRC).

The SIMC guidelines are also intended to be used in conjunction with other more specific agreements or arrangements, including both existing agreements between parties to the NHIA and agreements with regard to the subsequent release to third parties of data owned by jurisdictions which are parties to the NHIA. For example, the NHMRC's *National Statement on Ethical Conduct in Research Involving Humans* (which was issued in 1999 and is currently under review) requires all epidemiological research using de-identified data to be approved by a Human Research Ethics Committee (HREC). It is also the practice of some State health authorities to require external users of patient level health data to sign 'conditions of release' covering specific confidentiality requirements such as the purpose for which data may be used, requirements for the storage and subsequent deletion of data, and restrictions on the publication of data or the provision of data to a third party. These

'conditions of release' may also cover other issues such as copyright, limitation of liability, professional indemnity and insurance.

Although many of the examples used to illustrate the guidelines focus on hospital data, the guidelines are intended to provide a number of principles which are broadly applicable to hospital morbidity data, data on community based health services and other health data. They are intended to be technologically neutral in the sense that they cover statistical data generated from health records regardless of whether these records are held in paper or electronic format. They are adopted by the SIMC as the basis of best practice and can be refined to meet local needs and circumstances.

The main focus of the guidelines is on the protection of the anonymity of individual patients. However, guidelines are also included outlining current practice in relation to the anonymity of public and private hospitals.

Enquiries about these guidelines should be referred in the first instance to the SIMC Secretariat.

Guidelines about the anonymity of individual patients

With regard to individual confidentiality, the overriding aim of these guidelines is to avoid the identification of individual persons in health data. Although health privacy legislation and policies vary between Australian jurisdictions, their common purpose is to govern the collection, use and disclosure of 'personal information' about the health of, or health services provided to, individuals whose identity is apparent or can be reasonably ascertained. 'Personal information' can be defined as:

> 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 be reasonably ascertained, from the information or opinion (whether directly from the information or from the information when read in combination with other information held by or available to the organisation).

This means that, before providing a health data set to other agencies, the providing agency must satisfy itself that either:

- if the data set is not anonymised, it will only be used or disclosed for purposes for which the use or disclosure of personal information is permitted by its policies and legislation, or
- the data set is anonymised in the sense that the identity of individual patients is not apparent, and cannot reasonably be ascertained, from the data set either on its own or in combination to any other information to which the user may have access.

The focus of these guidelines is on the second principle; their aim being to assist data custodians to manage the risk of identification of individual patients. Guidance on compliance with the first principle is provided by the *Guidelines under Section 95 of the Privacy Act 1988* issued by the NHMRC. Further guidance on whether or not a given use of data needs to be approved by an HREC is also provided by the NHMRC's *National Statement on Ethical Conduct in Research Involving Humans* (issued in 1999 and currently under review) and its document *When does quality assurance in health care require independent ethics review?* (endorsed on 20 February 2003).

There are two types of risk that need to be managed when health data are provided for statistical use.

The first risk is that a person who is the subject of the data set may be identified, even if no information other than the fact that he or she is in the data set is disclosed. There is a risk that information may be disclosed at a level of detail that would enable a user to deduce that an individual had been admitted to hospital or was a client of some other health service. This risk may arise, for example, when there are only one or two people resident in a small community with a given combination of five year age group, sex and country of birth. If these data items are released then it may be possible for an individual to be identified as a person who has used a health service during the reference period. The disclosure of this fact alone may raise privacy issues for the person concerned.

The second risk is that, if a person is identifiable (because of information already in the public domain, or known privately by a data user), further information may be disclosed. For example, if a user of statistical data knows that a particular individual has been a patient of a health service (because that information is in the public domain, or known privately by the user) and could potentially identify that individual in the data set, there is a risk that the user may be able to ascertain further information such as the patient's diagnoses or treatment. For example, this type of risk may arise in small communities where it may be common knowledge that a person with specific demographic characteristics had been admitted to hospital. However, a similar risk may arise when data are released for larger areas; for example, even at State level there may be only one or two patients with certain rare diagnoses and the identity of the patient(s) may be widely known from unusual circumstances surrounding the case, and the disclosure of other diagnoses or procedures may constitute a breach of privacy.

Although these risks may be of particular concern with regard to hospital data, similar risks arise in relation to other health data. A list of some of the data items which pose particular risks of identification, drawn from the various National Minimum Data Sets managed by the SIMC, is given in Appendix 1.

Application of the guidelines about the anonymity of individual patients

Privacy legislation in Australian jurisdictions generally allows identifiable health information to be used for some purposes such as research projects which have been approved by a Human Research Ethics Committee; these guidelines are not intended to restrict such uses of information. However, it is good practice for data custodians to negotiate data requirements with researchers to ensure that the level of detail provided does not exceed that which is reasonably required for the research project.

Similarly, these guidelines are not intended to apply to the use of patient level data for statistical linkage with other data collections. In fact data linkage between two collections can only occur if the data from each collection are sufficiently detailed to uniquely distinguish between different patients. However, the guidelines can be used as an indicator of good practice when decisions are made concerning the release of data after the linkage has taken place.

Guideline 1:

Ensure that only those data items essential to the user's purpose are released.

It is not good practice to provide more information than is needed for a specific project. In fact, a requirement that users consider carefully which data items and what degree of specificity are needed will assist in complying with Guideline 2 below. For example, users should consider at an early stage of the project whether they need age rather than date of birth (and at what level of detail – years, months, etc) and whether they need a simple metropolitan/rural split rather than postcode or Statistical Local Area (SLA) of residence.

Guideline 2:

Ensure that the pool of people who could potentially have contributed to a cell (the denominator population) is as large as possible while still enabling the user to do their job. This can be achieved by aggregating domain values. Specifically, patient level data should not be provided with a combination of demographic data items that distinguishes groups with an estimated population of less than 1,000.

This guideline, sometimes known as the '1,000 denominator population rule' has been developed by the Australian Institute of Health and Welfare (AIHW) as good practice for the release of data. It is an example of the 'data reduction' principle described in Appendix 4 of the National Statistical Service Handbook. The populations are defined using any demographic information that is relevant (in the sense that the combinations of the demographic data items may enable individuals in the community to be identified) and for which resident population estimates are available from the Australian Bureau of Statistics (ABS). Hence the populations may be defined on the basis of geography (e.g. postcode or region of residence), age, sex, country of birth, Indigenous status or marital status. (Similar risks which may arise in relation to more specific data items such as dates of birth, admission or separation are addressed by Guidelines 3 and 4.)

Thus, for example, data which are disaggregated by five year age group and Statistical Local Area (SLA) of residence will not be provided if the estimated population in some five year age groups resident in some SLAs is less than 1,000. To overcome this, some age groups and/or SLAs may be combined into broader age groups or larger geographic regions. This will usually be done in consultation with data users.

This rule is considered to be a useful guiding principle (although not perfect) because the 'denominator' population (i.e. the population in the community, not the 'numerator' population or the population of health service recipients) provides a measure of the risk of identification. At a practical level, because it is based on ABS population estimates rather than health service data, it can be applied to determine the level of aggregation that can be provided before the health service data are analysed, thus avoiding the need to analyse the data before a decision can be made as to the level of detail that can be provided.

However, there may also be situations where a more cautious approach is needed. For example, consideration may need to be given to other variables such as the location of a hospital to protect individuals from one area who may be admitted to a hospital in a different, distant area.

Guideline 3:

As a general rule date of birth should not be provided to users of health data (except, as stated above, for approved data linkage projects).

In the Australian context this guideline is a specific application of Guideline 2. There are about 36,500 dates in a century so, on average, the current population of Australia (about 20 million) would give an average denominator population of about 550 Australians with any given date of birth. Date of birth is commonly used as a linkage variable in statistical work but this type of work needs to take place using appropriate linkage protocols that protect confidentiality, and only in projects endorsed by ethics committees and data custodians. Provision of date of birth information in data that are to be used for purposes other than approved data linkage projects could give rise to an unacceptable risk that the data could be used for unapproved linkage that could result in identification of patients. This is recognised in the NHMRC Guidelines where the definition of 'identified samples or data' includes the statement that 'examples of identifiers may include the individual's name, date of birth or address'.

Most user requirements can be satisfied by age calculated from date of birth and the date of the service event. The level of precision to which the age is calculated may need to be negotiated with the user; while five year age groups will often suffice, some uses (e.g. those which focus on paediatric services) may require age in single years or even months or days. If age in days is a requirement of the research project, special arrangements may need to apply (such as undertakings by the data recipient) to minimise the risk that the information gives rise to the identification of patients.

Guideline 4:

Caution should be adopted in relation to the provision of data items that pose a high risk of identification because they may be used to identify particular health service events, such as dates of admission to, or separation from hospital, or in some cases long length of stay and low or high birth weights.

Again this is a specific application of Guideline 2. It is often a combination of data items rather than a single data item that poses a risk of identification.

All jurisdictions consider dates of admission and separation to pose a high risk of identification in the same way as date of birth. However, this issue requires a common sense approach rather than a strict rule. For example, many small rural hospitals admit less than 100 patients per annum and the date of admission alone may enable many of these patients to be identified. In a larger hospital this would not usually pose a risk, but there may be some admitted patients with an unusually long length of stay who could be identified from the combination of the dates of admission and discharge. Similarly in a large maternity hospital there may be a small number of deliveries with unusually low birth weight and this may enable the baby and its mother to be identified.

For most analysis purposes, month of admission, month of separation and length of stay information is sufficient. If actual dates are required then additional precautions may be required to minimise the risk of identifying individual patients, such as limiting the detail provided by other variables.

Although as a general rule the AIHW does not provide dates of health service events, it has on occasion made an exception where the denominator population is large (i.e. more than 10,000 rather than 1,000) and there is a need for information on the actual dates.

Guideline 5:

As a possible approach to maintaining the anonymity of individual patients in statistical tables derived from health service data, cells showing less than five (5) health service events may be suppressed or aggregated unless exceptions are agreed between national and State/Territory data custodians.

This data suppression rule, or a slight variation of it, is currently applied by three State/Territory health authorities (Victoria, Tasmania and the NT) when considering requests for hospital morbidity data. It is an example of the data suppression rules described in Appendix 4 of the National Statistical Service Handbook. Its main value is as a rule of thumb that may assist in the identification of some cells in statistical tables or data cubes with potential identification problems. For example, it may assist in identifying cells where the diagnosis or procedure (or some other non-demographic characteristic) may risk identification. However, it does not assist with identification of cells that relate to small denominator populations.

The SIMC is reluctant to endorse this guideline as a general rule because, although some jurisdictions use it, it is not considered essential by all jurisdictions. There is, however, some benefit if those jurisdictions which do apply this data suppression rule agree on a particular (albeit slightly arbitrary) number such as five separations. It is generally more important to consider the size of the pool of people who could have potentially contributed to the cell in question (the denominator population) rather than the number of cases in the cell (the numerator population). It should also be noted that a cell showing five health service events need not necessarily represent five people because individuals may have multiple health service events (e.g. as many as 150 hospital separations per annum for a single dialysis patient). The blanket application of this type of rule can also create technical problems for large interactive data products where cells with larger numbers of health service events may need to be suppressed in order to prevent the generation of cells with small numbers.

On balance, the SIMC considers small cell sizes to be a useful indicator of the likelihood that some statistical table entries or some other statistical product may generate a risk that individual patients may be identifiable. When cells with less than five health service events occur it may be worth considering aggregating variables in order to provide better anonymity – for example, presentation of age in 10 year rather than 5 year age groups. As a general rule this would be preferable to the suppression of cells which may often contain important information. However, the user may prefer more detail in a table with some suppression as opposed to less detail in a table with no suppressions; for example, the fact that there were less than five health service events in a particular category may still be useful information.

Guidelines about the anonymity of individual hospitals

Guideline 6*:*

As a basic approach to maintaining the anonymity of individual private hospitals or private hospital owners in statistical tables derived from hospital morbidity data, cells should be suppressed if:

- *there are fewer that three (3) reporting units, or*
- *there are three or more reporting units and one contributed more than 85% of the total separations, or*
- *there are three or more reporting units and two contributed more than 90% of the total separations.*

This is a modification of a rule adopted some years ago by the AIHW and State/Territory health authorities in order to address the 'commercial in confidence' nature of private hospital data.

Guideline 7:

Decisions about the release of data identifying individual public hospitals should be referred to the relevant State/Territory health authority for consideration.

This guideline is included here for completeness. The SIMC has noted that the provision of information about the performance of individual public hospitals is the prerogative of State/Territory health authorities and does not consider it appropriate to offer guidance on this matter. It should be noted, however, that the appendices to *Australian Hospital Statistics*, available on the AIHW web site, do make available some broad information such as bed numbers on individual public hospitals.

Appendix

NMDS data items which pose a particular risk of client identification

1. Demographic data items

The following demographic items, which are common to a number of National Minimum Data Sets (NMDSs), pose a particular risk that their provision in combination and/or without some degree of aggregation may enable particular individuals (e.g. members of a small geographic or cultural community) to be identified as patients of a health service.

Area of usual residence

Country of birth

Age

Date of birth

Establishment identifier (particularly for establishments with small catchment areas)

Indigenous status

Marital status

Person identifier

Preferred language

Sex

2. Health service data items

The following data items, included in NMDSs covering particular types of health service event, 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. In some cases the risk may relate mainly to outliers (i.e. those health service events for which the data item has an unusual value) while in other cases the risk may arise from an unusual combination of values.

2.1 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)

2.2 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)

2.3 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

2.4 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

2.5 Community mental health care NMDS

Mental health legal status

Principal diagnosis

Service contact date

2.6 Elective surgery waiting times NMDS

Indicator procedure

Listing date for care

Reason for removal from elective surgery waiting list

Date of removal

2.7 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.)

2.8 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.)

2.9 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)