HealthStats NSW

Privacy issues and the reporting of small numbers

September 2015

More reports from HealthStats NSW
are available at:
www.healthstats.nsw.gov.au

NSW GOVERNMENT | Health

NSW MINISTRY OF HEALTH
73 Miller Street
NORTH SYDNEY NSW 2060
Tel. (02) 9391 9000
Fax. (02) 9391 9101
TTY (02) 9391 9900
www.health.nsw.gov.au

SHPN (CEE) 150486

ISBN 978 1 76000 281 7

Suggested citation: Centre for Epidemiology and Evidence. 2015. HealthStats NSW: Privacy issues and the reporting of small numbers. Sydney: NSW Ministry of Health.

Further copies of this document can be downloaded from the HealthStats NSW website at www.healthstats.nsw.gov.au

Produced by: Centre for Epidemiology and Evidence, Population and Public Health Division, NSW Ministry of Health.

Email: healthstats@moh.health.nsw.gov.au

September 2015

Contents

Acknowledgements

This report was originally prepared in 2011 by Dr James Scandol and Helen Moore from the Centre for Epidemiology and Evidence, with revisions in 2015 by Sarah Neill. The authors wish to thank Elise McHugh, Victoria Monahan and Charlotte Roberts for their detailed comments on the legal sections, and Dr Lee Taylor for suggestions which improved the report.

1 Introduction

HealthStats NSW is an interactive, web-based application that allows users to access data from a variety of different sources and tailor reports about the health of the New South Wales (NSW) population, for their own use. HealthStats NSW provides information on:

- the health status of the NSW population
- health inequalities and the determinants of health
- the major causes of disease and injury and current health challenges
- trends in health and comparisons between population groups (by age, sex, ethnicity, rurality and socioeconomic status) and geographic locations.

The electronic publishing of health indicators and health data are associated with requirements to ensure compliance with privacy laws and NSW Health privacy policies. Privacy refers to the right of an individual to have their personal health information safeguarded from loss, misuse and unauthorised access and disclosure. Privacy is of the utmost importance and NSW Health is required to ensure the protection of the privacy of individuals and/or communities whose health information may be reported on HealthStats NSW. Publication of health statistics is also associated with professional responsibilities for epidemiologists and biostatisticians to communicate the inferential limitations of that data, which can be particularly constrained when small numbers are reported.

This report presents an overview of the privacy laws and policies relevant to HealthStats NSW, and then summarises the key components of statistical disclosure, and the methods developed to reduce the risk of such disclosures. Statistical issues associated with the reporting of small numbers are also discussed.

The final section of the report examines the architecture of HealthStats NSW and outlines how the Centre for Epidemiology and Evidence manages the indicators presented in the application.

This combination of system architecture and the indicator configuration aims to minimise the risks of statistical disclosure of both individuals and small Aboriginal communities. The majority of indicators presented in HealthStats NSW also include confidence intervals of calculated statistics so that the inferential limitations are apparent.

The overall approach taken to manage privacy issues in HealthStats NSW includes:

- ensuring there is a public interest in any information published
- ensuring information that may identify an individual or community is not published
- following state and national guidelines concerning the appropriate number of people reported in both the numerator and denominator of calculated statistics
- using statistical disclosure control methods to manage privacy risks and small number issues in HealthStats NSW.

2 Background

Government agencies collect and use data on individuals and organisations to inform both day-to-day decisions, and the development and implementation of longer-term policies. Implementation of evidence-based policies requires the collection and management of data. Development of performance indicators from those data supports our understanding of whether particular policies are achieving their goals. Thus data collection programs, data management systems and public performance reporting systems are a key component of government transparency and accountability.

These activities are carried out within a legal and ethical context with respect to the privacy of individuals. The rapid increase in the use of information technology by all sectors of society has seen a rising awareness about the potential to compromise individuals' privacy. Various policies and laws have been developed in NSW and elsewhere to ensure that such technological advances do not pose risks to the privacy of individuals.

NSW Health and the Local Health Districts manage large volumes of information, some of which includes confidential data about individuals and communities. These data are used, among other things, to infer trends in the health of the population of NSW, illustrate inequalities in health that exist within the state, and indicate the effectiveness of various policies and programs.

There is thus a balance to be achieved in the publication of population health information. Providing too little information (or information that has been too highly aggregated) may limit the ability of government to make informed decisions on particular issues and publicly report on the effectiveness of programs aimed at improving the health of the community. In contrast, providing information in too much detail may compromise the privacy of individuals or communities. With the development of powerful web-based reporting systems such as HealthStats NSW, there is ongoing interest in understanding this balance and developing robust strategies to ensure that both private and public interests are met.

Many of the statistical strategies to protect privacy in reporting systems are relatively simple, although some require very technical algorithms. Some of these strategies will be presented in this report, but the value-based nature of many disclosure and privacy issues must be recognised. Most of these values will be captured in contemporary privacy laws and NSW Ministry of Health policies, but some aspects or interpretations of disclosure and privacy are of an ethical and professional nature.

3 Legislative, policy and ethical context

3.1 Privacy laws and policy in NSW

In NSW, the publishing of health information is subject to privacy laws and legislation. Privacy obligations arise primarily from two separate statutes, however there are other relevant pieces of legislation which impose specific controls on when and how information can be used and disclosed in NSW.

Privacy and Personal Information Protection Act 1998 (NSW)

The *Privacy and Personal Information Protection Act 1998* regulates non-health personal information in the public sector in NSW (e.g. employee records).

Health Records and Information Privacy Act 2002 (NSW)

The *Health Records and Information Privacy Act 2002*, or HRIP Act, regulates personal health information in the public and private sectors in NSW. The health system relies upon the principles contained within the HRIP Act to use and disclose personal health information.

The NSW Privacy Commissioner has published statutory guidelines under the HRIP Act to clarify the interpretation of certain elements of this Act. There are two statutory guidelines relevant to HealthStats NSW: statutory guidelines on research (Privacy NSW 2004a); and the statutory guidelines on the management of health services (Privacy NSW 2004b).

As stated in the HRIP Act (Part 1, Section 6), de-identified information is exempt from privacy law and from the requirements of the NSW Health *Privacy Manual for Health Information* (Section 5.2 – Personal Information). This report therefore aims to provide further detail regarding the disclosure control methods utilised in HealthStats NSW to protect the privacy of individuals and communities in NSW.

NSW Health *Privacy Manual for Health Information* 2015

The NSW Privacy Commissioner notes that the HRIP Act statutory guidelines should also be read along with the NSW Health *Privacy Manual for Health Information* (NSW Health 2015). This manual provides operational guidance to the legislative obligations imposed by the HRIP Act. The manual outlines procedures to support compliance with the HRIP Act in any activity that involves personal health information.

It is important to note that the HRIP Act and the statutory guidelines only apply to identifiable information. For information to be classified as de-identified it must not contain identifiers which, if linked with other information, could lead to the identity of a person. If there is a "reasonable" chance that the information is potentially identifiable, it cannot be classified as de-identified and is therefore not exempt from privacy law obligations.

Government Information (Public Access) Act 2009 (NSW)

A complementary legislative instrument to the HRIP Act is the NSW *Government Information (Public Access) Act 2009* (GIPA Act). The GIPA Act is designed to facilitate open and transparent government by encouraging the proactive public release of government information by agencies.

Decisions with regard to the release of information under the GIPA Act must be put to the public interest test (s13) which balances the public interest for disclosure with the public interest against disclosure.

The GIPA Act does not, therefore, provide any authority to override privacy law in NSW, but continues to emphasise the subtle balance between the public interest of access to Government-held information and the privacy interests of NSW citizens.

Public Health Act 2010 (NSW)

The other significant legislative instrument in NSW that authorises the collection of health information is the *Public Health Act 2010*. Section 130 prevents the release of personal health information unless certain criteria are met. For example, Section 83 places a requirement upon the Chief Executive of a hospital to notify the Secretary of Health about patients who have (or have had) a notifiable disease (as defined in Schedule 2). Disclosure conditions are also associated with such information (s56, Protection of Identity).

Privacy Act 1988 (Commonwealth)

The *Privacy Act 1988* regulates the Commonwealth public sector and the private sector. As this includes non-government organisations and private sector health providers in NSW, it is useful for NSW Health to be aware that this Act may impact on the way these organisations choose to share personal health information with NSW Health.

National Health Information Standards and Statistics Committee

The National Health Information Agreement (Australian Institute of Health and Welfare 2013) supports the development of more consistent policies across Australia for statistical reporting. The National Health Information Standards and Statistics Committee was formed in 2008 and assumed the roles previously undertaken by the Statistical Information Management Committee, the Health Data Standards Committee and some of the roles of the National Health Performance Committee.

Prior to the formation of the National Health Information Standards and Statistics Committee, the Statistical Information Management Committee presented *Guidelines for the Use and Disclosure of Health Data for Statistical Purposes* (SIMC 2007). This document includes seven reporting guidelines to increase the anonymity of individual patients in hospital-based health statistics. These guidelines are still current and are presented in more detail later in this report.

3.2 Regulation of the reporting of de-identified information

HealthStats NSW reports on de-identified health information. Care has been taken to ensure that published data cannot be linked (or joined) to other data, available publicly or within other organisations. This is to ensure that any data from HealthStats NSW cannot be re-identified. If there is any reasonable chance that data has the potential of being re-identified, disclosure control methods are applied to ensure that the privacy of the individuals or communities is maintained. The key phrase in both the HRIP Act and the statutory guidelines is "reasonable steps to de-identify", which is explained in the statutory guidelines (Privacy NSW 2004ab, p8) as:

> *"When de-identifying information, you should consider the capacity of the person or organisation receiving the information to re-identify it or re-link it to identifiable information. Removing the name and address may not always be enough, particularly if there are unusual features in the case, a small population, or there is a discussion of a rare clinical condition. Reasonable steps to de-identify might also include removing other features, such as date of birth, ethnic background, and diagnosis that could otherwise allow an individual to be identified in certain circumstances."*

3.3 Publication of Aboriginal health statistics

The NSW Ministry of Health and the Aboriginal Health and Medical Research Council (AHMRC), the peak body representing Aboriginal Community Controlled Medical Services in NSW, have developed the *NSW Aboriginal Health Information Guidelines* (NSW Health 1998). These Guidelines provide a framework of ethical and culturally sensitive protocols for the collection and use of personal health information relating to Aboriginal and Torres Strait Islander peoples in NSW.

The purpose of the Guidelines is to ensure consistency and good practice in the management of health and health-related information about Aboriginal people in NSW. There are 11 guiding principles underlying the Guidelines which indicate that health-related information should be used to support improved health and better planning and delivery of health services (Principle 2). An additional principle is that the utilisation rather than the collection of information be maximised (Principle 9). Principle 1, however, states that the management of health and health-related information about Aboriginal peoples must be ethical, meaningful and useful to Aboriginal peoples.

The Guidelines also expand the usual interpretation of de-identified information to include:

> *"de-identified information – information which has been stripped of details such as individual names, addresses, dates of birth, death or other events, or in certain circumstances Aboriginal community identifiers; or where such details have been sufficiently altered to render the identification of individuals or communities unlikely. (There are cases where aggregated data, apparently stripped of identifiers, may permit individuals to be identified, e.g. an uncommon medical condition. Special consideration should be given to ensuring the privacy of individuals and communities in such circumstances.)"*

Note that these Guidelines extend the notion of individual privacy to community privacy. Indeed, the Guidelines include the definition:

> *"privacy (of Aboriginal community information) – the right of an Aboriginal community to exercise appropriate control over the availability of Aboriginal community information to others."*

The AHMRC have developed *Guidelines for Research into Aboriginal Health*. These Guidelines detail the criteria under which the AHMRC Ethics Committee reviews and approves research projects. The document states that:

> *"The ethics committee will only approve a project where there is net benefit for Aboriginal people and communities."*

The same principles apply to the data published on HealthStats NSW. The public health utility of the data and associated net benefit for the community must be considered as greater than the risk to those communities whose data is published.

These issues also have a national research context. The Australian National Health and Medical Research Council (NHMRC), in collaboration with the Australian Research Council and the Australian Vice-Chancellors' Committee, has published a *National Statement on Ethical Conduct in Human Research* (NHMRC 2014) which includes several references to privacy, including Guideline 1.11:

> *"Researchers and their institutions should respect the privacy, confidentiality and cultural sensitivities of the participants and, where relevant, of their communities".*

This Guideline clearly indicates that privacy and confidentiality issues also have a community dimension. In the NHMRC publication *Values and Ethics: Guidelines for Ethical Conduct in Aboriginal and Torres Strait Islander Health Research* (NHMRC 2003), there is a more detailed reference to community privacy in National Statement 1.19:

> *"Where personal information about research participants or a collectivity is collected, stored, accessed, used, or disposed of, a researcher must strive to ensure that the privacy, confidentiality and cultural sensitivities of the participants and/or collectivity are respected."*

3.4 Professional ethics associated with the publication of small numbers

There are also issues associated with the professional ethics of epidemiologists and biostatisticians which can constrain the publication of small numbers in data tables. Apart from the privacy issues raised above, there are also obligations within these professions not to publish information that could result in unreliable statistical inferences. As Steel et al. (2003) discuss, small numbers are subject to much larger relative variance which makes any inferences drawn from these numbers much less reliable. If statistical errors are more likely, then the justification for publishing these numbers is compromised.

If there are valid reasons for needing this information, then the records would be available within government or available to interested persons after they had agreed to non-disclosure policies and were also aware of the inferential limitations of that data. There are thus professional responsibilities on biostatisticians to consider the value of publishing tables with small numbers, particularly if the uncertainties associated with any inferences drawn are not presented in parallel. For example, the International Statistical Institute Ethical Principle 8 states "... statisticians should ensure that they accurately and correctly describe their results, including the explanatory power of their data. It is incumbent upon statisticians to alert potential users of the results to the limits of their reliability and applicability." Similar guidelines exist for the American Statistical Society.

The Statistical Society of Australia Incorporated (SSAI) includes rules of professional conduct in its code of conduct. This includes a similar but more general rule concerning the duties of members to act in "the public interest" including that "Members are encouraged to advance public knowledge and understanding of statistics and to counter false or misleading statements." There is also a rule relating to the maintenance of ethical standards including that "... members shall ensure that the collection of information and the publication of results shall observe relevant privacy laws" (SSAI 2015).

4 Statistical disclosure issues

The literature of statistical disclosure recognises several types of disclosure which require consideration. Each type of disclosure has a distinct interpretation with respect to the reasoning behind potential constraints to public reporting, and their relationship to privacy are discussed below (original citation Dalenius, 1977, but used by Duncan et al. 1993).

4.1 Identity disclosure

Identity disclosure occurs when an individual is able to be identified by information that is publicly released. This will obviously occur if someone's name, address or photograph is published, but could also occur if that person could be identified by linking the released information with other information that was available. The chance that this could occur will rise substantially if that person is from a small or identifiable community. For example, inadvertent identity disclosure could occur if a statistical table indicates one person in a small community had a rare disease, and it was known in that community that a particular person had spent a lot of time in hospital that year with a rare disease (Navarro 2008). However, the most likely cause of identity dislosure will be the inadvertent or deliberate release of the name, address, photograph or geocoded location of an individual.

The purpose of de-identifying (or anonymising) data is to dramatically reduce the risk of identity disclosure (but the risk will not be eliminated completely with unit-level records, where each row in a data table represents information on a particular person). Once the data have been de-identified, these data are not considered confidential and are not subject to privacy law. There are, however, other disclosure considerations associated with de-identified data.

4.2 Attribute disclosure

Attribute disclosure occurs when a characteristic about a person is released. The actual person is not identified (that would be defined as identity disclosure), but attribute disclosure could, via a chain of events, result in identity disclosure. An additional concern with attribute disclosure is that, under certain circumstances, more information than was known about a person may be revealed.

Say, for example, that it is commonly known that one person in a small town has a rare disease such as listeriosis. If statistical tables are published that disclose that the individual with listeriosis also has syphilis, then clearly privacy laws have been breached.

Reducing the risk of attribute disclosure is the primary focus of statistical disclosure control (see Section 6). If identity disclosure results from attribute disclosure, then the application of privacy law will become relevant. However, attribute disclosure of an individual (even without identity disclosure) is still considered an issue of concern for the following reasons: information about an individual has been released (even if we cannot identify the individual); the public interest arguments (in relation to the epidemiological utility of the data) associated with publishing any information about an individual (or a small number of individuals) are questionable; finally, the statistical inferences drawn from such a small sample are likely to be very compromised.

4.3 Inferential disclosure

Inferential disclosure is more subtle, and ultimately of less concern, than either identity or attribute disclosure. Inferential disclosure occurs when, using published information, characteristics about an individual can be inferred from statistical patterns or models. For example, if provided with all the independent variables, a regression model can be used to predict the dependant variable for a particular person. Such a situation is unlikely and the predicted value will not likely be a precise or accurate estimate of the observed variable if the model has been developed for statistical purposes. Inferential disclosure will not be further considered in this report.

4.4 Community disclosure

An additional type of disclosure which has not been formally defined in the statistical literature is associated with groups of people rather than individuals, and is of particular relevance in reference to the reporting of information about Aboriginal people. Given the concept of "community privacy" described above, there must be an associated concept of "community disclosure". Community disclosure will have parallels to individual disclosure, with the same issues associated with unique identifiers and attributes, but in this case the appropriate response is likely to involve processes (such as consultation with the community involved, including some form of community consent) as well as outcomes (such as restructured tables).

Figure 1 demonstrates the relationship between the risk of statistical disclosure, the associated epidemiological utility of the data, and the degree of aggregation of the data. There is often a trade-off in the collection and reporting of population-scale health data which must be considered. In most cases, providing too little data (or data which have been too highly aggregated) may compromise the epidemiological utility of the data to the end-user. However, providing too much data (or data which is not aggregated sufficiently), may compromise the privacy of individuals or communities. Note there is diminishing marginal epidemiological utility to providing less aggregated data for the reasons discussed above but the disclosure risks will increase rapidly.

Clearly, there are situations where individually identifiable data is necessary for public health practitioners (e.g. contact tracing for an infectious disease outbreak) but in such scenarios access to such information is very tightly controlled. The challenge for HealthStats NSW is to determine a threshold of acceptable risk of attribute disclosure when reporting public health data, which is the primary focus of statistical disclosure control for HealthStats NSW.

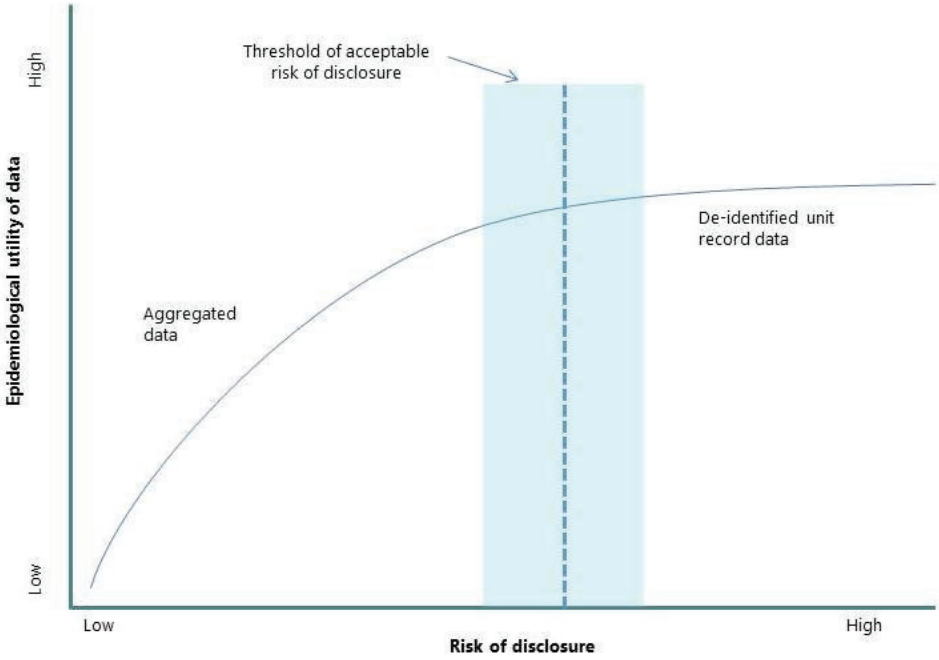


Figure 1. *Association between the risks of disclosure and the epidemiological utility of data, for data aggregated at varying levels (ranging from highly aggregated data to de-identified unit record data). Note that the threshold of acceptable risk will not be clear cut and will depend upon the context.*

5 Managing privacy issues and small numbers in HealthStats NSW

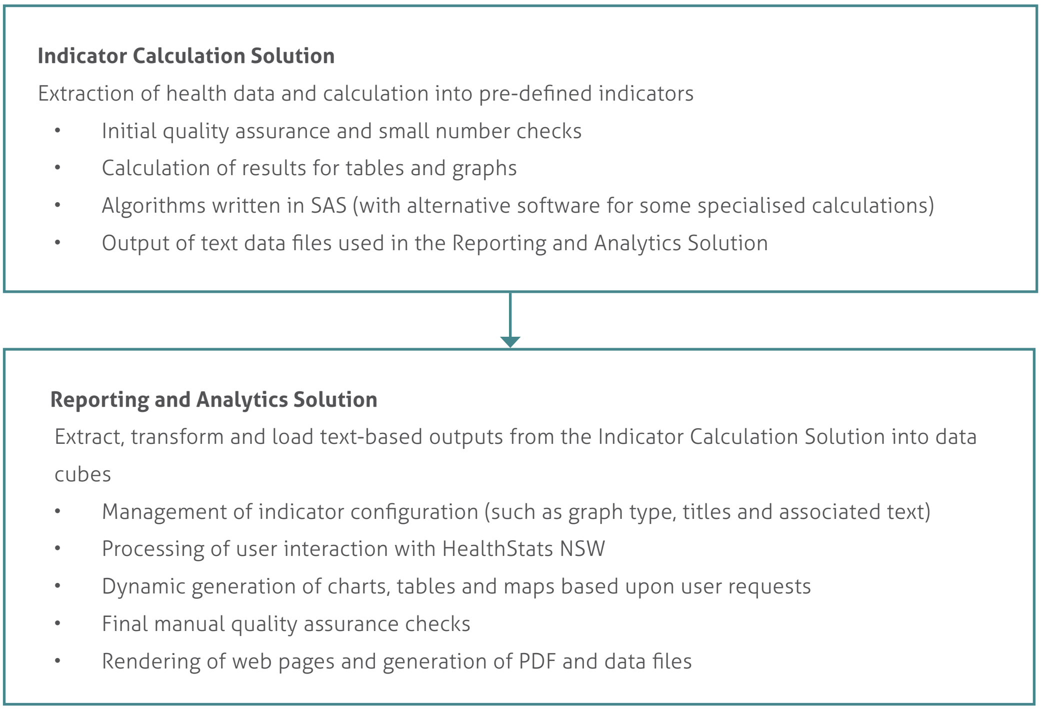
5.1 HealthStats NSW system architecture

There are many approaches which may be used in the development of web-based reporting systems, the outputs of which generally fall into two categories:

- Pre-compiled reports: Users search for, and then have delivered, a pre-compiled report of an indicator or highly aggregated dataset in tabular form (normally cross-tabulated
- Online query system: Users step through a screen-by-screen workflow to define a query and then execute that query. The results are usually returned in a cross-tabulated form (and sometimes a chart).

The architecture of HealthStats NSW provides an excellent basis for minimising the risks associated with statistical disclosure. This is because the system is designed primarily to be a report-delivery system rather than a web-based data query tool. In HealthStats NSW, users will perceive they have a very large number of choices, but are actually constrained in a similar fashion as to how they would be if they were searching for pre-compiled reports. This strategy results in a major decrease in disclosure risk because the final outputs are highly controlled.

The underlying technical architecture of HealthStats NSW is complex and the details are beyond the scope of this report. There is, however, value in describing the two major components of the system as they pertain to privacy and small number issues. Figure 2 illustrates the relationship between the Indicator Calculation Solution (an application based on SAS™ software) and the Reporting and Analytics Solution (implemented with Microsoft Business Intelligence Tools). The two solutions are separated by appropriate security technologies.



Indicator Calculation Solution

Extraction of health data and calculation into pre-defined indicators

- Initial quality assurance and small number checks
- Calculation of results for tables and graphs
- Algorithms written in SAS (with alternative software for some specialised calculations)
- Output of text data files used in the Reporting and Analytics Solution

Reporting and Analytics Solution

Extract, transform and load text-based outputs from the Indicator Calculation Solution into data cubes

- Management of indicator configuration (such as graph type, titles and associated text)
- Processing of user interaction with HealthStats NSW
- Dynamic generation of charts, tables and maps based upon user requests
- Final manual quality assurance checks
- Rendering of web pages and generation of PDF and data files

Figure 2. *Schematic representation of the two major components of HealthStats NSW system architecture*

Currently in HealthStats NSW, the extraction and transformation process, and data quality and integrity process, occurs in the Indicator Calculation Solution (in SAS). This system is composed of a set of algorithms applied in SAS to process unit-level or semi-aggregated data from the population health data warehouse into health indicators. By undertaking these initial quality assurance and small number checks in the Indicator Calculation Solution, any data with potential privacy issues are tightly secured by NSW Health information technology systems.

The public-facing aspect of HealthStats NSW is the Reporting and Analytics Solution. This solution imports SAS text-based outputs of aggregated data (counts and rates), creates data cubes and renders charts, tables, maps and PDF reports on request. The integration between the two solutions involves the passing of aggregated data between the solutions. A final manual quality assurance check is undertaken by an appropriate staff member working on the Reporting and Analytics Solution prior to each new data release. This check is to certify the accuracy of the data and ensure consistence with the relevant privacy policies and guidelines.

Quality assurance checks may be automatically applied in the Indicator Calculation Solution. This would ensure that statistical disclosure control methods are applied in a routine and consistent manner, and all relevant disclosure risks have been addressed prior to the aggregated data being passed between solutions, thereby ensuring the maximum privacy of individuals and communities.

6 Statistical disclosure control

6.1 HealthStats NSW disclosure control risk assessment process

An assessment of the disclosure control risks associated with each indicator report on HealthStats NSW is undertaken prior to each new data release, as illustrated in Figure 3.

A key component of the HealthStats NSW disclosure control risk assessment process involves the identification of particular attributes that have the potential to raise privacy concerns for each indicator report. The key attributes which are considered in the HealthStats NSW disclosure control risk assessment process include those identified in Table 1.

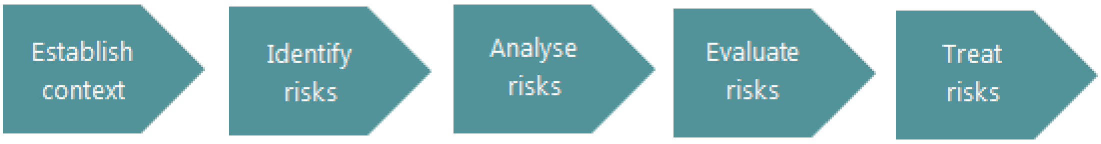


Figure 3. *HealthStats NSW disclosure control risk assessment process*

Table 1. *Summary of attributes associated with privacy risks*

Attribute	**Privacy concerns**
Small area	• Small areas often report small numbers of people (lower denominators) that increase the risk of identity disclosure. • Small areas are also likely to contain small communities which may present additional community disclosure risks.
Aboriginality	There are additional policy obligations associated with the reporting of Aboriginal health information relating to both individual and community disclosure.
Infectious disease	There is a social stigma attached to many infectious diseases which are likely to present unique privacy concerns.
Analysis type	Different methods of analysis present variable risks. The reporting of count data carries a high risk, whereas calculated values such as life expectancy have small disclosure risks.
Small numbers	As the number reported (*n*) increases, the risk of attribute disclosure decreases. As a general rule, reporting cell counts *n<5* is not recommended (in order to preserve privacy of individuals), although this depends on the size of the denominator.
Census or survey	• Reports based upon stratified random surveys have inherently less risk to privacy than reports based upon a census. • In addition, the type of data reported may impact privacy risks: • individual/case-based (e.g. deaths or births): carries a higher risk as there is more chance of an individual being identified • service-based (e.g. hospitalisations): carries a lower risk, as one hospitalisation does not necessarily reflect one individual. *For example, an individual undergoing dialysis may be counted multiple times in hospitalisation data, thereby making it difficult to determine the number of individuals who were actually admitted for a particular condition.*

The relationship between data attributes and disclosure risk can be defined in a flexible and diverse manner. For example, the attributes in Table 1 can be mapped to the magnitude of disclosure risk in a variety of ways. Figure 4 illustrates examples of such patterns with the allocation of either a high or low risk score dependent on the data reported in each indicator report.

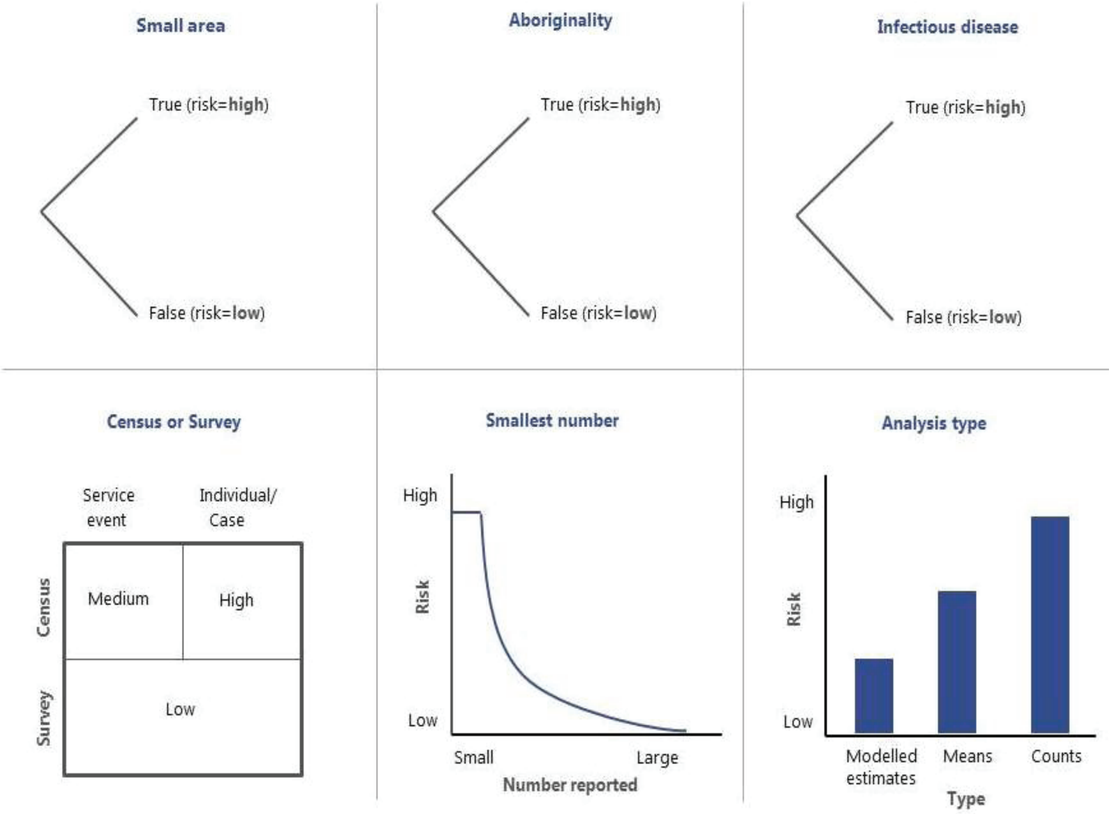


Figure 4. *Assigning privacy risk scores for indicator attributes in HealthStats NSW*

Results from the mappings of data attributes to potential privacy risk can be combined to provide information regarding the magnitude of disclosure risks, from which an assessment can be made in relation to particular indicators on HealthStats NSW. Indicators with larger risks are reviewed and restructured as required to manage the risks.

6.2 Disclosure control methods

There has been significant research on statistical disclosure and there exists a toolbox of strategies from the field of statistical disclosure control which have varying costs, risks and benefits. Two types of information are commonly presented in statistical tables: count (frequency) data or magnitude data. These two types of data present different challenges for statistical disclosure control and will be considered separately.

Rules for count data

Count data present the number of individuals who meet certain categorical criteria. These criteria are usually specified by the intersection of a row and column, and perhaps page, which define a particular cell in a table. In some cases, count data are reported as relative frequencies or percentages of a category or combination of categories.

The US Federal Committee on Statistical Methodology (2005) identified two general approaches for disclosure control of count data: special rules and the threshold rule.

- Special rules are simply agency-specific conditions used to constrain the resolution of a public report. For example, ages will never be presented at a finer scale than a 5-year age class. Tables are then constructed based upon these rules.
- The threshold rule is more generic and is usually stated that table cells with less

than *k* individuals (sometimes referred to as sensitive cells) must not be reported. Common values of *k* are 3, 5 or 10. If a table cell reports a number less than *k*, then there are a number of general strategies for managing the disclosure risk associated with that table including: table restructure; cell suppression; and changing table values. These are discussed below.

Table restructure

The first, and simplest, is to restructure the table to increase the number of people reported in that cell by combining rows or columns until the threshold *k* is reached. Sometimes this will require human judgment about what groups should be sensibly combined and the process may be difficult to automate for some types of tables, particularly if there is not a natural hierarchy or order in the classification used.

Cell suppression

The second commonly used approach is cell suppression: replacing the contents of a table cell with a number below a threshold value with an identifiable character such as an asterisk (*).

Cell suppression is a commonly used approach but has associated complications. For example, if marginal totals are provided (i.e. total counts for a row or column) then complementary suppression of another table cell may be required. Otherwise readers will be able to calculate the count in the suppressed cell using subtraction. In large tables, the patterns of complementary suppression required may become exceedingly complex and specialised algorithms in linear programming need to be used to determine the effective patterns of cell suppression. This problem can become even less tractable if, for example, the marginal totals have been presented in another table in the same or a different report. Elliot (2001) discusses the complexities associated with these so-called table linkages and the complementary concept of table differences.

It should be noted that, while the cell suppression rule is easy to implement, it may be unnecessary in relation to disclosure risk if the small number is drawn from a large population (see the "Denominator rule" below). Conversely, a larger number of cases (above the specified threshold) for a sensitive issue drawn from a smaller population may represent an unacceptable risk of disclosure.

Changing table values

The third approach, changing table values, usually involves some type of rounding or randomisation of cells values. Rounding (say to the nearest five people) requires care as it usually results in marginal totals that are incorrect, which can erode public confidence in the table. More sophisticated algorithms such as controlled rounding or controlled tabular adjustment can be applied to maintain accurate marginal totals but these require the implementation of more complex algorithms to fully implement.

Randomisation involves replacing the contents of a table cell with a number which has been randomly selected below a threshold value. This approach is commonly used by the Australian Bureau of Statistics. For example, when reporting suicide deaths in children (5–17 years), cell counts with small values have been randomly assigned to protect the confidentiality of individuals. In this example, readers are informed in the table key that randomisation has been applied.

One aspect of some health indicators which does not seem to have gained much attention is that in some cases the health records in a cell will be based upon the same person. For example, if a person is re-admitted to hospital within a year there may be two hospital separations recorded, but they are for the same person. Unless a unique identifier is available for a person (such as will be available for a linked dataset), and a query which processes only unique cases is used, then many tables will *overestimate* the number of individuals being reported from administrative datasets. In such cases, more conservative threshold rules may need to be used.

Denominator rule

As discussed above, in relation to small areas or sub-populations, the size of the population denominator may be more relevant to disclosure than the size of the numerator. The Australian Statistical Information Management Committee (SIMC) guidelines (SIMC 2007) suggest that statistical results involving small numbers can be presented if the population from which they are drawn is more than 1000 people. The reasoning behind this is that, even for rare conditions, in populations or communities larger than 1000 people, the likelihood of identifying an individual would be very small.

It should be noted that an emphasis within the SIMC guidelines is on the denominator, rather than the numerator, associated with statistics. The SIMC argues that risks to privacy are more commonly associated with size and nature of the population that statistics are drawn from rather than the actual number of people reported.

Rules for magnitude data

Magnitude data are based upon a simple statistic of a numerical characteristic of individuals represented by a table cell. This simple statistic is usually the sum (such as the total income), but it could also be the average or a percentile. Magnitude data require additional considerations for statistical disclosure control beyond that required for count data (Federal Committee on Statistical Methodology 2005).

Tables of magnitude data are usually tested against the (*n,k*) dominance rule, which is a generalisation of the threshold rule. The (*n,k*) dominance rule checks if *n* unique entities contribute more than *k*% of the value of that cell. If a cell breaks the (*n,k*) dominance rule, then the same types of approaches as described above are used to modify the table until all cells are considered safe from disclosure.

HealthStats NSW indicators which are based on magnitude data are generally either calculated from large Australian Bureau of Statistics surveys or a census of births and deaths, therefore none of these tables represents any significant risk of disclosure.

6.3 Managing disclosure risks in HealthStats NSW

The following provides an overview of the methods currently utilised by HealthStats NSW to manage disclosure control risks in indicator reports. Whilst these are the methods currently employed, there are a number of additional methods which may also be considered and utilised, if deemed appropriate and in keeping with the risk assessment process described in Section 6.1. These may include methods such as attribute aggregation.

Justification of the indicator presented based upon health priorities and policies in NSW

Indicators presented in HealthStats NSW have been developed on the basis of the need for that information from planners and policy makers. There must be a "public interest" argument for all indicators. For example, the *NSW State Plan 2021* (NSW Government 2011) includes a number of indicators of Aboriginal health which can be supported by systems such as HealthStats NSW without compromising either individual or community privacy by developing indicators that are not based upon small populations or rare conditions. The number and emphasis of indicators presented in HealthStats NSW that are associated with Aboriginal health reflects the importance of these policies in NSW.

System architecture that does not store unit-level data on public-facing servers

Section 5.1 outlined the system architecture for HealthStats NSW. A key aspect of this approach is that no unit-level data are stored on public-facing servers. Ministry officers requiring access to unit-level data (which are still de-identified) are required to comply with strict security and confidentiality policies, and these servers are not public-facing.

Consideration of high-risk attributes associated with privacy and disclosure risks

Prior to releasing new indicator reports on HealthStats NSW, various attributes associated with privacy and disclosure risks have been considered (see Section 6.1). This includes privacy risks associated with the reporting of information about Aboriginal people, small numbers, infectious disease, and small areas, whilst taking the analysis type and data source into consideration. These potentially high-risk attributes are assessed prior to applying appropriate disclosure control methods (see Section 6.2).

Design tables to minimise the number of cells with denominators less than 1000 people or counts of individuals in table cells which are less than five people

As indicated above, the Australian Statistical Information Management Committee (SIMC) presented guidelines (SIMC 2007) which suggest that statistical results should not be presented if the population from which they are drawn is less than 1000 people. The vast majority of populations from which statistics are calculated in HealthStats NSW are greater than 1000 people. For example, hospitalisations by Local Health District are in the order of 8000–140 000 separations per year per District. Acknowledging that some of these counts are for the same individual (which cannot be enumerated without using linked datasets), these populations still provide for very substantive denominators for any hospitalisation statistics based upon NSW hospital admissions data.

The SIMC guidelines do not fully endorse the application of a specific data suppression rule because of the inconsistency with which these rules are applied in Australia and the determination by some jurisdictions that such a rule is not essential (SIMC 2007). When applied, the SIMC suggests a minimum value of five individuals within a table cell. They also note that table redesign is usually the preferred strategy to increase counts in cells, rather than removing the data and marking cells with an asterisk (or other character).

As a response to this issue, some indicators in HealthStats NSW have tables redesigned to increase the counts being reported. For example, aggregation of ARIA categories 'remote' and 'very remote' is commonly used to increase counts in these categories as some conditions are particularly rare. Aggregation of 5-year age groups into 10-year age groups is also required for some indicators.

Note that HealthStats NSW does contain some tables with small cell counts. These indicators have been assessed on a case-by-case basis and deemed to have a low risk of disclosure because, applying the denominator rule, the population from which the observations are drawn is very large (e.g. the population of males in a Local Health District).

Statistical smoothing of results for small populations

When indicators are required for small populations such as Local Government Areas, identifiable subsets of these populations, such as Aboriginal people, will sometimes have counts that are less than 1000 people. In these cases, Bayesian statistical smoothing methods are used to model the patterns in the data (Lawson et al. 2003) and the estimated, rather than actual, numbers are reported. This approach masks the counts of individual people but retains the key statistical inferences that can be drawn from the data. Simpler smoothing methods such as moving averages are also sometimes used when presenting some time-series data.

Quality assurance processes that track privacy concerns

The Centre for Epidemiology and Evidence uses a robust quality assurance system to record the development of indicators for HealthStats NSW. This system is used to monitor the broad components that constitute an indicator. At present, the components include the indicator data, indicator graphs, data tables, associated text content and any privacy issues associated with the indicator.

This quality assurance system aims to ensure that privacy issues do not become "lost" during the complex process of indicator configuration. The software used (Atlassian 2015) tracks commentary about privacy issues and also describes the strategies that have been used to minimise disclosure risks.

Responsiveness to privacy concerns

HealthStats NSW includes a highly visible privacy page that is accessible from anywhere in the application. The page summarises privacy issues associated with HealthStats NSW and includes a link to this report and to the NSW Health *Privacy Manual for Health Information*. Should a person feel their privacy has been compromised by any results presented on HealthStats NSW, they are welcome to contact the Centre for Epidemiology and Evidence and outline their concerns. If required, particular indicators can be taken offline until these issues are resolved. If a person believes their privacy has been breached they may also contact the Senior Privacy Officer, NSW Ministry of Health, or the Information and Privacy Commission NSW.

7 Discussion

Privacy and small number issues represent an important challenge for web-based reporting systems. Indicators are a key component of the performance management, planning and evaluation required for health systems, and therefore play an important role in government accountability and transparency. However, the privacy of the individuals whose personal experiences constitute these health statistics cannot be compromised, and the interpretive limitations of these data must be communicated.

Indicators presented in HealthStats NSW have been developed on the basis of a need for the information from planners and policy makers. However, ethical and legal issues must also be considered and addressed prior to the publishing of indicators on HealthStats NSW. Epidemiologists and biostatisticians have a professional ethical obligation to not publish information which could result in unreliable statistical inferences as small numbers are subject to much larger variance which makes any inference drawn from these numbers much less reliable. This issue must be considered together with the privacy issues relating to the publishing of small numbers as, in addition to compromising the privacy of individuals, there is an increased likelihood of statistical errors associated with the reporting of small numbers. When assessing the epidemiological utility of the data, limitations on both the reporting and interpretation of the data must be considered. Fortunately, there are well-established laws, policies, guidelines and methods that enable web-based reporting systems such as HealthStats NSW to present health indicators without compromising privacy or statistical inferences. The overall approach is to: ensure there is a public interest in any information published; ensure information that may identify an individual is not published; follow national guidelines concerning the minimum number of people in both the numerator and denominator of calculated statistics; and apply disclosure control methods to protect the privacy of individuals and communities when small numbers are inevitable.

Article 12 of the Universal Declaration of Human Rights states that "No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence …", and there is widespread acceptance of the concept of individual privacy in Australian society. Health statistics, which are so crucial for the evidence-based provision of health services, are built-up from the records of private citizens, all of whom place their trust in the health services to maintain their rights to privacy.

References

Atlassian (2015) JIRA – Issue and Project Tracking Software, Sydney. Available from: www.atlassian.com/software/jira. Accessed 15 July 2015.

Australian Institute of Health and Welfare (2013) National Health Information Agreement. Available from: http://meteor.aihw.gov.au/content/item.phtml?itemId=583436&nodeId=file53be175f402ec&fn=NHIA_2013.pdf. Accessed 14 May 2015.

Duncan GT, Jabine TB & de Wolf VA (eds). (1993) Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics. National Research Council, National Academy Press, Washington, 274 pp.

Elliot M (2001) Disclosure Risk Assessment. In: P Doyle, JI Lane, JJ Theeuwes & LV Zayatz (eds). Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies (pp. 75-90), Elsevier, London.

Federal Committee on Statistical Methodology (2005) Statistical Policy Working Paper 22 (2nd version). Washington: Office of Management and Budget, 128 pp.

Lawson AB, Browne WJ & Rodeiro CL (2003) Disease Mapping with WinBUGS and MLwiN, John Wiley & Sons, Chichester.

Navarro R (2008) An ethical framework for sharing patient data without consent. Informatics in Primary Care 16: 257–62.

NHMRC (2003) Values and Ethics: Guidelines for Ethical Conduct in Aboriginal and Torres Strait Islander Health Research. National Health and Medical Research Council. Commonwealth of Australia, Canberra, 24 pp.

NHMRC (2014) National Statement on Ethical Conduct in Human Research 2007 (Updated May 2015). The National Health and Medical Research Council, Australian Research Council and the Australian Vice-Chancellors' Committee Australian Government, Canberra, 107 pp.

NSW Government (2011) NSW 2021: A plan to make NSW number one. NSW Government. Available from: https://www.nsw.gov.au/sites/default/files/nsw_2021_plan.pdf. Accessed 9 July 2015.

NSW Health (1998) NSW Aboriginal Health Information Guidelines, State Health Publication AHB980128, 11 pp.

NSW Health (2007) Aboriginal Health Impact Statement and Guidelines, PD2007_082, North Sydney, 17 pp.

NSW Health (2011) Health Outcomes Information Statistical Toolkit, NSW Health, Sydney.

NSW Health (2015) Privacy Manual for Health Information, North Sydney, 75 pp.

Privacy NSW (2004a) Statutory guidelines on research. Health Records and Information Privacy Act 2002, 32 pp.

Privacy NSW (2004b) Statutory guidelines on the management of health services. Health Records and Information Privacy Act 2002, 16 pp.

SIMC (2007) Guidelines for the Use and Disclosure of Health Data for Statistical Purposes. Statistical Information Management Committee, Australian Institute of Health and Welfare, 16 pp.

Steel D, Green J & Brown L (2003) Best Practice in Small Area Analysis and Reporting – Literature Review and Guidelines. Centre for Health Service Development, University of Wollongong, 108 pp.

www.healthstats.nsw.gov.au