national STATISTICS

Health and care 2006

Review of the Dissemination of Health Statistics: Confidentiality Guidance

Working Paper 2: Risk Assessment

Office for National Statistics
1 Drummond Gate
London SW1V 2QQ
Tel: 020 7533 9233
Fax: 020 7533 9292

Contact points

For enquiries about this publication contact:

Statistical disclosure centre on 0845 601 3034

Email: info@statistics.gsi.gov.uk

For general enquiries, contact the National Statistics Customer Contact Centre:

Tel: 0845 601 3034

Minicom: 01633 812399

Email: info@statistics.gsi.gov.uk

Fax: 01633 652747

Post: Room 1015, Government Buildings, Cardiff Road, Newport NP10 8XG

About the Office for National Statistics

The Office for National Statistics (ONS) is the government agency responsible for compiling, analysing and disseminating economic, social and demographic statistics about the United Kingdom. It also administers the statutory registration of births, marriages and deaths in England and Wales. The Director of ONS is also the National Statistician and the Registrar General for England and Wales.

A National Statistics publication

National Statistics are produced to professional standards set out in the National Statistics Code of Practice. They are produced free from any political influence.

Contents

1 Introduction 1

2 Disclosive situations 2

Misreporting units 2

3 Statistical units 3

4 Linking or combining tables 4

References 7

1 Introduction

This working paper provides more information on issues concerned with risk management to support the guidance for the review of the dissemination of health statistics.

In order to be explicit about disclosure risks in tables of health statistics, a risk assessment should be undertaken by considering a range of disclosive situations. The situations considered are used to identify 'unsafe' cells within the table and determine a minimum level of protection.

The risk may be heightened if:

- any other disclosive situations are likely to occur
- statistical units are represented more than once in the table
- groups of statistical units are represented in the table
- tables based on the dataset have already been released
- other freely available datasets can be linked to the tables

2 Disclosive situations

Three disclosive situations are outlined in the consultation document as a guide. In many cases not all of these situations will be appropriate, in other cases additional situations may need to be considered. In order to identify the risks a range of disclosive situations should be considered for the different types of statistical units that require protection.

Misreporting units

Another disclosive situation that could occur is when units misreport. Someone knows information about a statistical unit but is unable to find that unit in a table and thereby discloses some additional information.

Disclosure would arise in this situation if the relevant cell contained a zero. In order to protect against this scenario all non-structural zeros within the table are considered unsafe. A zero is structural when a count in a cell is impossible. A structural zero is non-disclosive since one would know for certain that the cell was zero and hence will not have disclosed any further information. In general this scenario can be discounted due to the fact that protection cannot be provided for units that misreport information.

Example

A woman told her husband that she had a termination on certain grounds, but a table of health statistics contains a zero in this cell and therefore reveals that this did not occur.

3 Statistical units

Whether defined as individuals, households or businesses, the statistical units represented in the dataset that require protection should be identified. A particular issue that arises in health statistics is the protection of practitioners, GP practices, NHS staff, etc. In some cases the identity of the practitioner, etc will need to be protected but in other cases the statistics by their nature necessarily reveal their identity (eg statistics about the performance of a NHS Trust).

Care needs to be taken when statistical units are represented more than once in a table, for example if protection is required for practitioners then cells in a table where all the values (which could be large) relate to a particular practitioner could potentially be disclosive.

Disclosure risks may also increase if groups of statistical units (eg patients from a particular clinic) are represented in a table and therefore could identify each other.

Examples

Abortion statistics provide an example of a dataset where a number of different units are represented and the complex relationships between the units will affect confidentiality protection. The abortion (or the event) is directly related to an individual woman and the details of the female involved in the abortion need to be protected. Protection of these details requires the protection of the identity of the household of which the female who had the abortion is a member. In addition, an abortion is an event that can be directly related to a practitioner. Therefore, the identity of the practitioner, as the person who carried out the procedure, needs to be protected. Similarly the identity of the hospital/clinic as the place in which the termination occurred should also be protected.

Annual hospital admissions for particular conditions provide an example of a dataset where a statistical unit can be represented more than once in a table. The statistical unit is the individual patient who may have more than one admission within a year. This will need to be considered when defining which parts of the output pose a disclosure risk.

4 Linking or combining tables

Linking together different tables produced from the same dataset (that could have already been released) or other freely available datasets should not lead to disclosure.

When linked tables are produced from the same dataset it is not sufficient to consider the protection for each table separately. If a cell requires protection in one table then it will require protection in all tables, otherwise the protection in the first table could be undone.

Where tables provide data in terms of rates or percentages the numbers themselves may not be disclosive. However, if the rate or percentage is based on an unsafe cell and it is possible (by linking with other tables) to recover the original count then the cell with the rate or percentage is itself unsafe. Some protection can be provided by rounding rates or percentages. However, care still needs to be taken to avoid disclosure. Protection will be provided if the base from which the rate or percentage is calculated is sufficiently large since the implied count could be a range of values, however, this range must be large enough to satisfy disclosure rules and thresholds.

Another problem that can occur with multiple or linked tables produced from the same data source is called disclosure by differencing. This problem occurs when two or more tables, taken together, enable by subtraction or deduction the value for a potentially disclosive cell to be revealed.

Table 1

Age	<18	18-45	>45
Count	5	26	13

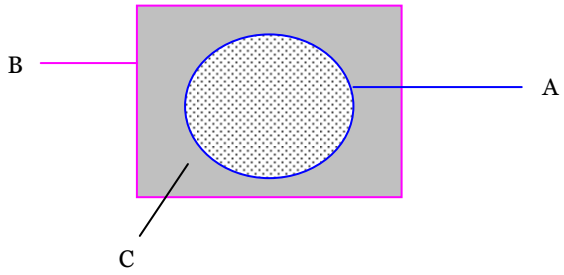
Table 2

Age	<19	19-45	>45
Count	6	25	13

Suppose Table 1 and Table 2 generated from the same dataset provide counts for a particular characteristic by age group. One can easily calculate the count for 18 year olds by differencing the count for the age band 19-45 and 18-45. The counts for the age band 18-45 and 19-45 maybe considered safe but their difference reveals a small and therefore potentially disclosive count for 18 year olds.

Suppose tables are produced for the two geographical areas A and B shown figure 1, where area A is a subset of B. One could easily produce a table for the shaded geographical area C. If two tables are produced for different geographies from the same data set then disclosure by differencing can occur even if the two tables have been protected independently.

Figure 1



Disclosure by differencing occurs when one grouping of the data is contained within another, eg 18-45 year olds is a subset of the range 19-45, and area A is a subset of area B. The two groups can be differenced to produce a residual area, eg 18 year olds or area C, for which statistics may be below the confidentiality threshold and therefore disclosive. The general overlapping of two groups does not necessarily create a disclosure problem, eg if a table displays counts for 18-45 year olds and a second table displays counts for 30-50 year olds, differencing the two would not reveal a disclosive count.

Differencing may be possible when aggregations of areas are considered. If a table displays counts for the age ranges 16-45 and 46-50 and a second table displays counts for 18-20 and 21-50 year olds, differencing could reveal a disclosive count for 16 and 17 year olds. In some cases differencing will not result in small compact areas or groups, in fact for geographical differencing this may often be the case.

Any disclosure control method that does not perturb or modify large numbers within the table will not protect against disclosure by differencing. For example rounding will provide protection against disclosure by differencing but suppression will not. In this case two options are available, either restrict the format for releasing the data or evaluate all potential differences.

Restricting the format for releasing data, eg only releasing data on specific coterminous geographies and using standard variable categories is simple to implement but can be very restrictive since different users require statistics in different formats. Restricting the format for releasing data may be appropriate if the use of the data is limited and can be standardised. For data that are required by a range of users working with different categorisation systems or geographies then this solution is not appropriate, although the risk is minimised the utility will be significantly compromised.

Another approach to this problem is to evaluate all differences to assess whether disclosure problems are likely to arise. This is likely to be a complex task and in many cases would need to be automated, for example in the case of disclosure by differencing from geographical boundaries Geographical Information Systems (GIS) may need to be employed.

The evaluation would need to be carried out at the smallest geographical level or the most detailed categorisation available. In addition the evaluation would need to be repeated for different datasets. If a large number of different tables are to be created from the same microdata file using a different categorisation and/or different geographies then the task to evaluate and protect against disclosure by differencing would be very complex and time consuming. However, if only a limited number of tables were planned for release, for example on two different geographical boundaries and the GIS software and expertise required was available then this solution could be feasible.

More information on disclosure by differencing can be obtained from Brown (2003) and Duke-Williams and Rees (1998).

References

Brown, D (2003) *Different approaches to disclosure control problems associated with geography*, ONS, United Nations Statistical Commission and Economic Commission for Europe Conference of European Statisticians, Working Paper No. 14.

Duke-Williams, O and Rees, P (1998) 'Can Census Offices publish statistics for more than one small area geography? An analysis of the differencing problem in statistical disclosure', *International Journal of Geographical Information Science* 12, 579-605.