Health and care 2006

Review of the Dissemination of Health Statistics: Confidentiality Guidance

Working Paper 3: Risk Management

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

© Crown copyright 2006

Published with the permission of the Controller of Her Majesty's Stationery Office (HMSO)

You may re-use this publication (excluding logos) free of charge, in any format for research or private study or internal circulation within an organisation. You must re-use it accurately and not use it in a misleading context. The material must be acknowledged as crown copyright and you must give the title of the source publication. Where we have identified any third party copyright material you will need to obtain permission from the copyright holders concerned.

For any other use of this material please apply for a free Click-Use Licence for core material at: www.opsi.gov.uk/click-use

or write to: Office for Public Sector Information, St Clements House, 2-16 Colegate,
Norwich NR3 1BQ.

Fax: 01603 723000

Email: hmsolicensing@cabinet-office.x.gsi.gov.uk

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 Table redesign 2

3 Suppression 4

Examples 4

4 Rounding 6

5 Random rounding 7

6 Controlled rounding 8

7 Barnardisation 9

8 Record swapping 10

9 Alternative methods for presenting data 11

10 Overview of methods 14

References 16

List of tables

Table 1: Age and type 4

Table 2: Age and type, primary suppression 5

Table 3: Age and type, primary and secondary suppressions 5

Table 4: Age and type, random rounding, base 5 7

Table 5: Age and type, controlled rounding, base 5 8

Table 6: Age and type, barnardisation 9

Table 7: Age and type, percentages 11

Table 8: Age and type, rates per 1,000 11

Table 9: Redesign 14

Table 10: Suppression....14

Table 11: Controlled rounding 14

Table 12: Barnardisation 14

Table 13: Record swapping 15

List of figures

Figure 1: Age and type, count 12

Figure 2: Age and type, rate 13

1 Introduction

This working paper provides more technical information on disclosure control methods to support the guidance for the review of the dissemination of health statistics. These methods can be used to manage the risk of disclosure in tables of health statistics by disguising those cells identified as being unsafe. More details can be found in Willenborg and de Waal (1996) and Doyle et al (2001).

2 Table redesign

Three examples of table redesign are:

- grouping or collapsing categories within a table
- aggregating to a higher level geography or for a larger population sub-group
- aggregating tables across a number of years/quarters/months

The advantage for disguising unsafe cells using table design is that the original counts in the data are not damaged. However, the detail displayed within the table will be reduced. The method is easy to implement but does require a good knowledge of the data and an awareness of the needs of users in order to combine categories whilst maximising the utility in the data.

Examples of grouping categories within a table are:

- top or bottom coding – where values at the very top or bottom ends of the distributions of continuous variables are recoded into single categories (eg age under 15 years)
- broad-banding continuous variables so that a response is recoded to lie within a particular range of values (eg using age bands of 16–25 years, 26–35 years, etc)
- broad-banding or collapsing categorical variables so that they are grouped together into one category (eg combining similar medical procedures)

Categories with unsafe cells should be selected, and combined where possible with 'similar' categories. Also, two smaller similar categories might be combined to form a larger one, but if they are dissimilar, each should be combined with a different larger category to minimise the relative data damage. It is important to take into account how the proposed change will affect the consistency between tables and historic comparisons. Collapsing categories does not necessarily have to be implemented across a whole table but can be applied to sub-tables.

If tables contain too many unsafe cells, then one solution is to increase the frequency count for each cell by aggregating to a higher level of geography or for a larger population subgroup. This method is straightforward to implement. Apart from the loss of detail, the data need not be damaged: the published frequencies maintain their correct values. The risk of identification and disclosure is reduced since the individual frequencies will be larger and the population at risk is also increased.

A safe way to increase access to more detailed counts is to publish three (say) year aggregates. Not only does this increase the level of data that can be output, it also adds protection because the data is uncertain in timing (between 1 and 3 years). In addition it becomes much more difficult to make an identification due to the time lag and migration issues. The rules for defining unsafe cells are the same as those used for annual data.

An alternative to publishing aggregated data is to publish rolling aggregates, e.g. 2001+2002+2003 and then 2002+2003+2004, etc. If rolling aggregates are to be implemented then the rules for defining unsafe cells are not always straight forward and need careful consideration especially if any year of the rolling aggregate has been previously published.

3 Suppression

A method of protecting unsafe cells in tables is cell suppression. This means that unsafe cells are not published – they are suppressed and replaced by a special character, such as '..' or 'X', to indicate a suppressed value. This should be a different symbol from that used for missing values. Such suppressions are called primary suppressions. To make sure the primary suppressions cannot be derived by subtractions from published marginal totals, additional cells are selected for secondary suppression. The selection of secondary suppressions can be done either by hand or by software.

Where there are only a few unsafe cells in a table suppression will be a relatively easy method to implement and in most cases will not result in high information loss. A disadvantage of this method is that most of the information about the original values in the suppressed cells is removed and due to secondary suppressions some counts that are safe will also be removed. If the number of primary suppressions is not low then the information loss can be high and the ideal choice of secondary suppressions is not a trivial task.

A disadvantage of suppression is that this method does not offer a solution to disclosure by differencing. This would mean that without a detailed analysis of disclosure by differencing the statistics could not be published on any other geographies or other non-standard variable categories. A careful audit process would need to be implemented for any tables released on an ad-hoc basis, this could be time consuming and therefore resource intensive. Care will also need to be taken if suppression is used to protect linked tables. Any suppressed cells (primary or secondary) will need to be suppressed in all releases. Again this could result in a high level of checking.

Examples

Table 1 displays counts of treatment type 1 and 2 broken down by age bands. The shaded cell contains a frequency of 1 and is potentially disclosive.

Table 1: Age and type

	Age				
Outcome	< 12	12-15	16–19	> 19	Total
Type 1	1	5	7	6	19
Type 2	7	15	18	19	59
Total	8	20	25	25	78

If a threshold rule of less than 5 (< 5) were applied to Table 1 the frequency of 1 in the shaded cell could be replaced with an X, as in Table 2.

Table 2: Age and type, primary suppression

	Age				
Outcome	< 12	12-15	16–19	> 19	Total
Type 1	X	5	7	6	19
Type 2	7	15	18	19	59
Total	8	20	25	25	78

The suppressed cell in Table 2 can be derived by subtractions from the marginal totals. It is therefore necessary to carry out secondary suppressions.

Table 3: Age and type, primary and secondary suppressions

	Age				
Outcome	< 12	12-15	16–19	> 19	Total
Type 1	X	X	7	6	19
Type 2	X	X	18	19	59
Total	8	20	25	25	78

Table 3 shows the secondary suppressions (in the shaded cells) that are needed to ensure that the primary suppression is effective. Secondary suppressions should be chosen to minimise information loss, eg select internal cells before marginal totals and smaller counts before larger counts. Care should also be taken to ensure that suppressions are consistent throughout all releases. The process of secondary suppressions can become very complex as the number of suppressions increases. In order to ensure a safe and optimal solution, a disclosure control software tool (eg Tau-Argus, available at: http://neon.vb.cbs.nl/casc/) should be implemented.

4 Rounding

Rounding involves adjusting the values in all cells in a table to a specified base so as to create uncertainty about the real value for any cell while adding a small but acceptable amount of distortion to the data. Two alternative rounding methods are outlined below: random rounding and controlled rounding. In each case there is a choice of the base for rounding – common choices are 3 and 5. All rounded values (other than zeros) are then integer multiples of 3 or 5, respectively.

Although conventional rounding (where each cell is rounded to the nearest multiple of the base) does provide some protection it is not considered sufficient and is therefore not recommended here.

The advantage of using rounding is that if the number of unsafe cells is large then the table can be protected while still providing counts for all cells. Rounding will protect zeros without removing them since, within a table rounded to base 5, for example, a zero could represent any count between 0 and 4.

A disadvantage of rounding for protecting some health statistics is that there are difficulties in disguising cells in which the count of events can be associated with either 1 or 2 practitioners/hospitals whom it may be necessary to protect. For example, if a cell had an original count of 17 events all associated with one practitioner, then rounding this to 15 means that the count still relates to only one practitioner, the unsafe cell is not disguised. A further disadvantage of rounding is that some users require exact counts for particular statistics and rounded values would not be appropriate.

A major advantage of rounding is that it offers protection from disclosure by differencing since the difference between two rounded tables will also be rounded. This means that protection using rounding offers more flexibility in outputs compared with suppression. In order to fully protect against disclosure by differencing it may be necessary to increase the rounding base to a minimum of 5.

5 Random rounding

In random rounding, each cell value is rounded in a random manner, independently of other cells, usually (although not always) to an adjacent multiple of the rounding base. For example, values of 6, 7, 8 or 9 could be rounded to either 5 or 10, based on assigned probabilities. Various probability schemes are possible but an important characteristic is that they should be unbiased, ie there should be no net tendency to round up or down.

Table 4: Age and type, random rounding, base 5

	Age				
Outcome	< 12	12–15	16–19	> 19	Total
Type 1	0	5	5	5	20
Type 2	10	15	15	20	60
Total	5	20	25	25	75

Random rounding is relatively easy to implement. However, in some instances the protection can be unpicked. In order to ensure adequate protection, the resulting rounded table needs to be audited. After applying random rounding there may be inconsistencies in data within tables (rows or columns may not add up, eg row 1 does not sum to 20) and between tables (ie the same cell is rounded to a different number in different tables).

6 Controlled rounding

In controlled rounding, values in the cells of a table are rounded to a multiple of a common base in such a way as to preserve additivity to subtotals and table totals. The controlled rounding method works for hierarchical data and for linked tables. Table 5 shows a possible controlled rounding solution for Table 1. However, the method needs to be implemented in a statistical disclosure control software package, eg Tau-Argus (available at: http://neon.vb.cbs.nl/casc/).

Table 5: Age and type, controlled rounding, base 5

	Age				
Outcome	< 12	12–15	16–19	> 19	Total
Type 1	0	5	5	5	15
Type 2	5	15	20	20	60
Total	5	20	25	25	75

7 Barnardisation

Barnardisation is a post-tabular disclosure control method for frequency tables. The procedure modifies each internal cell of every table by +1, 0 or -1 according to the probabilities (p/2, 1-p, p/2). Zeros are unadjusted. The totals are added up from the perturbed internal cells. Typically, the probability p is quite small and therefore the majority of cells are not modified. As in most post-tabular adjustments it leaves inconsistent totals between tables. In addition, as in random record swapping, it leaves high risk in the small cells, ie the probability that a 1 is a true 1 is quite high. If the true value was a 1, then with probability p/2 the value is perturbed to a zero, with probability p/2 the value is perturbed to a 2 and probability (1-p) the 1 remains a 1.

Table 6 shows a possible solution when Table 1 is protected using barnardisation, with p=0.1. Many of the cell values have not been modified, including the 1; this may be considered unacceptable in terms of risk.

Table 6: Age and type, barnardisation

	Age				
Outcome	< 12	12–15	16–19	> 19	Total
Type 1	1	5	7	5	18
Type 2	7	16	18	19	60
Total	8	21	25	24	78

8 Record swapping

Pre tabulation techniques focus on perturbations of individual records using either a targeted or a random selection process. One such method is record swapping. This involves swapping characteristics between pairs of records that are partially matched (eg individuals that have the same age). Typically, in order to satisfy edit checks, swapping alters the geographic locations attached to the records, but leaves all other aspects unchanged. The effect on tabulations produced from the record-swapped data is that some of the data will be counted in the table for a different geographical location, depending on the level of geography chosen.

A potential disadvantage of implementing record swapping for certain health statistics is that a high level of swapping may be required in order to disguise all unsafe cells. The distribution of the statistics within tabulations produced from the record-swapped data would be distorted and a user would not be aware of the level or type of distortions.

An advantage of this method is that some uncertainty will surround any statistics derived through differencing two tables generated from the swapped records., Once sufficient perturbations have been carried out then any tables can be generated (using any categories for variables or any geographies) and the tables will be non-disclosive and protection will be provided from disclosure by differencing.

Record swapping was implemented for the US Census 2000 (see Zayatz 2003). A detailed discussion of alternative perturbation methods is provided by Brown (2003).

9 Alternative methods for presenting data

Alternative methods for presenting data can be considered as an approach to providing users access to information without disclosing the underlying data. In many cases this will provide a more robust analysis than reliance on the accuracy of small cell counts. This could include presenting data graphically or on a map or providing commentaries or analytical outputs. Care needs to be exercised to ensure that the outputs are safe. Any alternative method should not allow small counts to be identified. Some examples are provided.

Publishing rates, percentages, changes over time or indices may provide users with the information required without disclosing the underlying data. The protection provided by these methods will depend on how difficult it is to recover the underlying and potentially disclosive data. One must ensure that any implied counts satisfy confidentiality rules. 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 confidentiality rules.

Using the example above, in order to avoid publishing small counts the figures could be displayed as percentages (Table 7). This would be a useful way to present the data if users were interested in the age distribution of patients undergoing these treatment types. Row and column totals are not presented since although the percentages are rounded it would still be possible to work back to a small cell, eg 5 per cent of 19 is 0.95, which must therefore represent a 1.

Table 7: Age and type, percentages

	Age			
Outcome	< 12	12–15	16–19	> 19
Type 1	5%	26%	37%	32%
Type 2	12%	25%	31%	32%

If users were interested in the number of treatments relative to population size then this data could be displayed as rates (Table 8). As for percentages, if the population denominators are available then it may be possible to work back to the original counts.

Table 8: Age and type, rates per 1,000 (1 decimal place)

	Age			
Outcome	< 12	12–15	16–19	> 19
Type 1	1.7	19.8	4.7	0.9
Type 2	11.6	59.3	12.1	2.8

Graphs and maps are other alternative methods for presenting disclosive data and can be very useful for identifying trends and patterns. Again care needs to be exercised to ensure that the level of detail does not reveal unsafe data. Scatter plots should not allow the identification of outlying data points and maps should not allow individuals to be identified in a local area.

The map in Figure 1 displays counts of treatment for type 1 for under-12 year olds for wards in the Isle of Wight Primary Care Organisation. Note these data are fictitious and displayed as an example. The dots in the key represent a range of values, ensuring that the data are not disclosive.

Figure 1: Age and type, count

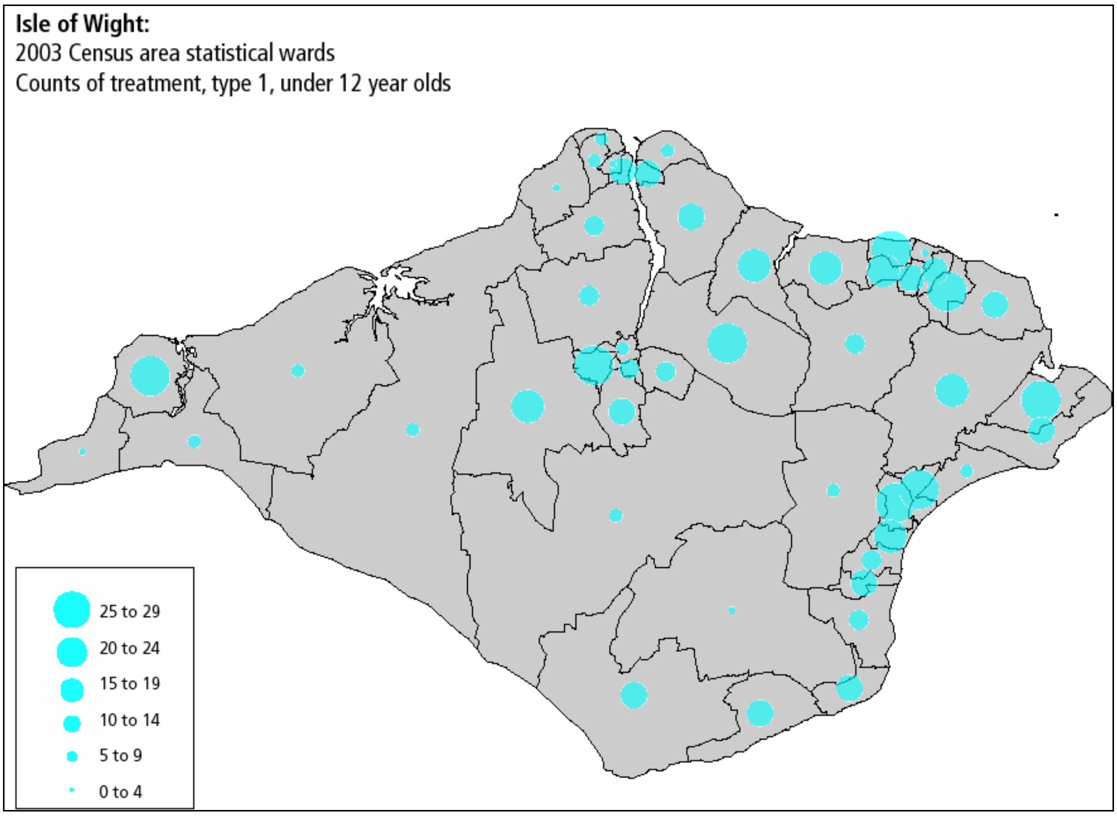
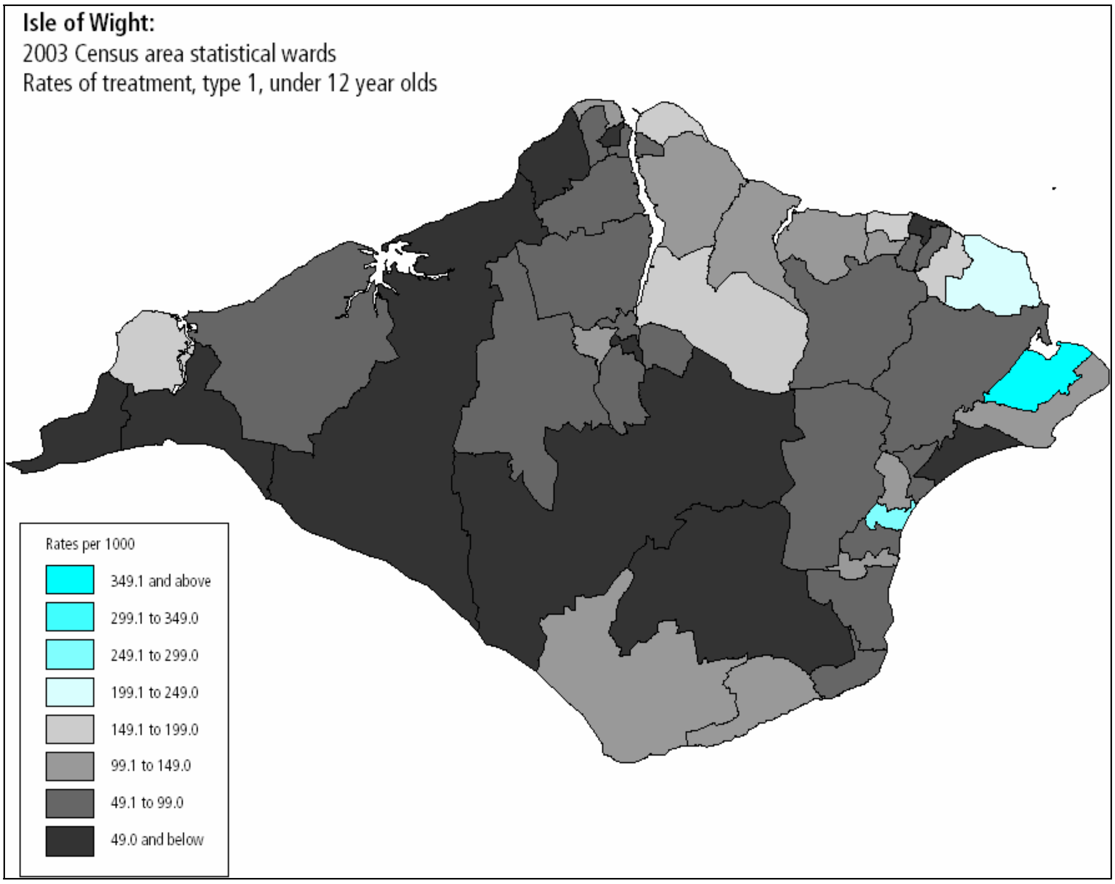

Isle of Wight:
2003 Census area statistical wards
Counts of treatment, type 1, under 12 year olds
25 to 29
20 to 24
15 to 19
10 to 14
5 to 9
0 to 4

Figure 2: Age and type, rate



The map in Figure 2 displays rates per 1,000 of the population. Again the key relates to a range of values. If population denominators are known then these ranges should be sufficiently wide to ensure that small counts cannot be identified.

10 Overview of methods

As an overview, the tables below display possible solutions to protecting Table 1 using different disclosure control methods.

Table 1: Age and type

	Age				
Outcome	< 12	12–15	16–19	> 19	Total
Type 1	1	5	7	6	19
Type 2	7	15	18	19	59
Total	8	20	25	25	78

Table 9: Redesign

	Age			
Outcome	< 15	16–19	> 19	Total
Type 1	6	7	6	19
Type 2	22	18	19	59
Total	28	25	25	78

Table 10: Suppression

	Age				
Outcome	< 12	12–15	16–19	> 19	Total
Type 1	X	X	7	6	19
Type 2	X	X	18	19	59
Total	8	20	25	25	78

Table 11: Controlled rounding

	Age				
Outcome	< 12	12–15	16–19	> 19	Total
Type 1	0	5	5	5	15
Type 2	5	15	20	20	60
Total	5	20	25	25	75

Table 12: Barnardisation

	Age				
Outcome	< 12	12–15	16–19	> 19	Total
Type 1	1	5	7	5	18
Type 2	7	16	18	19	60
Total	8	21	25	24	78

Table 13: Record swapping

	Age				
Outcome	< 12	12–15	16–19	> 19	Total
Type 1	1	5	7	5	18
Type 2	7	16	18	19	60
Total	8	21	25	24	78

References

Doyle, P, Lane, J I, Theeuwes, J J M and Zayatz, L (2001) *Confidentiality, Disclosure and Data Access: Theory and Practical Application for Statistical Agencies*. Elsevier Science BV: Amersterdam.

Willenborg, L and de Waal, T (1996) ‘Statistical Disclosure Control in Practice’, *Lecture Notes in Statistics* No. 111. Springer-Verlag: New York.

Zayatz, L (2003) *Disclosure Limitation for Census 2000 Tabular Data*, Bureau of the Census, United States, United Nations Statistical Commission and Economic Commission for Europe Conference of European Statisticians, Working Paper No. 15.