Health and care 2006

Review of the Dissemination of Health Statistics: Confidentiality Guidance

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

List of tables

List of figures and maps

Executive summary

This report represents the outcome of a comprehensive review of the dissemination of health statistics undertaken by the Office for National Statistics. It provides guidance for handling health statistics in a way that ensures the public interest in the figures is met while managing data confidentiality risks.

The reporting for this review has been a two-stage process. The first part of the review focused on developing guidance for published tables of abortion statistics (ONS 2005b). This report provides more general guidance on disclosure issues around published tables of health statistics derived from registration processes, administrative sources and statistical returns. Confidentiality issues concerned with microdata or record-level information are not covered by this review.

The guidance describes an approach that data providers should follow based on a general framework for addressing the question of confidentiality protection. The following elements of the framework are described in this document:

- Determining user requirements
- Understanding the key characteristics of the data and outputs
- Assessing disclosure risk by considering a range of potentially disclosive situations and identifying the parts of the table that could lead to disclosure. Producers of statistics are likely to find that outputs can be placed into one of three broad risk categories, defined in terms of the likelihood of an attempt to identify an individual and the impact of identification. Recommendations are made on the level of protection required for these three risk categories
- Legal and policy considerations
- Disclosure control methods. A number of different methods and techniques are presented and compared
- Implementation

More technical advice and worked examples are provided in associated working papers.

For many published tables, the risk of identifying individuals will be minimal and no disclosure control methods necessary. For other information the issues may be more complex. No single solution is available for these instances. Instead, guidance is provided on how to develop solutions for different types of datasets based on each of the steps of the framework. The guidance will allow data providers to apply appropriate solutions to confidentiality problems within their own organisation. This guidance replaces previous practices that have been adopted within the health field, such as the rule of thumb to suppress all values in tables less than 5.

1 Review and consultation on the dissemination of health statistics

Health statistics support a wide range of work to improve and protect our health, they inform patients and the public. Many of these areas of work require detailed figures, which may raise issues about data confidentiality. Producers of health statistics must ensure that their statistics meet the needs of users while at the same time protecting confidentiality.

This report results from a review of the dissemination of health statistics. The review was initiated in 2005 to address disclosure issues around published tables of statistics. Other forms of dissemination (for example, the provision of data to professional analysts) present different issues and will be addressed separately in a National Statistics guidance publication.

The aim of the review was to produce guidance for handling health statistics in a way that ensures the public interest in the figures is met while managing data confidentiality risks.

The review has been led by the Office for National Statistics (ONS) and has involved representatives from the Health Departments, Public Health Observatories and the devolved administrations. In addition the guidance has been released for public consultation.

The principles and approach outlined in this guidance will apply to all health statistics. However, this review and hence the examples, specific rules or guidelines presented are focused on tables derived from registration processes, administrative sources and statistical returns. These data sources have a complete coverage of the population or a sub-population. Specific guidelines for tables derived from sample surveys will be provided in a National Statistics publication to be produced by the end of 2006. Confidentiality issues concerned with microdata or record-level information are not covered by this review.

This review was established specifically for published health statistics, where following release there is no control over their further use. However, through consultation wider issues have been raised concerning data access and sharing that are recognised as important and need to be addressed. Guidance on the large scale transfer of data and confidential data for the production of statistics is available in the ONS publication, *Data Sharing for a Statistical Purpose: A Practitioners' Guide to the Legal Framework* (ONS 2005a). In addition a National Statistics publication will be produced by the end of 2006 to provide guidance on the provision of access to confidential data for specific persons for specific uses.

The guidance produced from the review is intended for anyone in the health community involved in the publication of health statistics. The Office for National Statistics (ONS) will implement the proposals and will advise ministers in Health Departments and devolved administrations to do the same. The guidance is also aimed at Primary Care Trusts, Public Health Observatories and other Arms' Length Bodies of the Health Departments.

This document describes the approach that data providers should take to meet users' needs while managing confidentiality risks. It proposes a general framework for addressing the question of confidentiality protection. The main elements of this framework are described in sections 4 to 9. These include understanding the data, assessing risk, and legal and ethical aspects. A number of possible methods for disclosure control are presented and guidance is provided on implementation. Technical advice is provided in the attached working papers:

- Confidentiality protection – legal and policy considerations
- Risk assessment
- Risk management
- Glossary
- References and other guidance

A worked example of this approach has been developed for abortion statistics (ONS 2005b).

2 Meeting users' needs while protecting confidentiality

> National Statistics will be valued for relevance, integrity, quality and accessibility – and produced in the interests of all citizens by protecting confidentiality. (National Statistics Code of Practice, Summary of Principles)

Health statistics provide an important public benefit. They are often of greatest value when they extend to small geographic areas or sub-groups. For example, it is well known that 1 in 3 people develop cancer at some point in their lives and 1 in 4 deaths are from cancer. We can conclude that cancer is a significant health risk, but little more. In order to examine health issues in more detail it is necessary to look at more detailed figures. A more informative view of the data could involve:

- analysing by age and gender to show the relative risks for these different groups
- breaking figures down by ethnic group and social class to reveal the extent of health inequalities
- examining local-area data to highlight problems of different localities
- investigating specific health information about proximity to potential health hazards such as mobile phone masts or landfill sites to allow an assessment of health risks

However, when statistics are released at a detailed level the risk of disclosing information about individuals is likely to be increased. Particular problems arise in tables containing small counts. For example table 1 shows counts of conceptions by usual place of residence and age of the mother at conception. (The data displayed are not true counts but are indicative of the true distributions.)

The table has many small counts, particularly for under 18s. Although the table does not itself reveal the identity of an individual there could be a risk that someone with a specific interest in this topic seeing a small number in the cell, could follow up private sources of information to locate the individuals and discover more details. There could also be a risk of disclosure from combining or linking this table with other information, eg a table of abortion or birth statistics.

Table 1: Counts of conceptions, by ward and age of mother

Ward	Under 18	Total
Ward A	5	56
Ward B	0	34
Ward C	3	94
Ward D	1	78
Ward E	2	66
Ward F	1	45
…..	…..	…..

The National Statistics Code of Practice and Protocol on Data Access and Confidentiality provide high level guidelines on how to approach such problems but there is a need for more detailed guidance on how to translate these policy statements into practice.

3 What is involved in confidentiality protection?

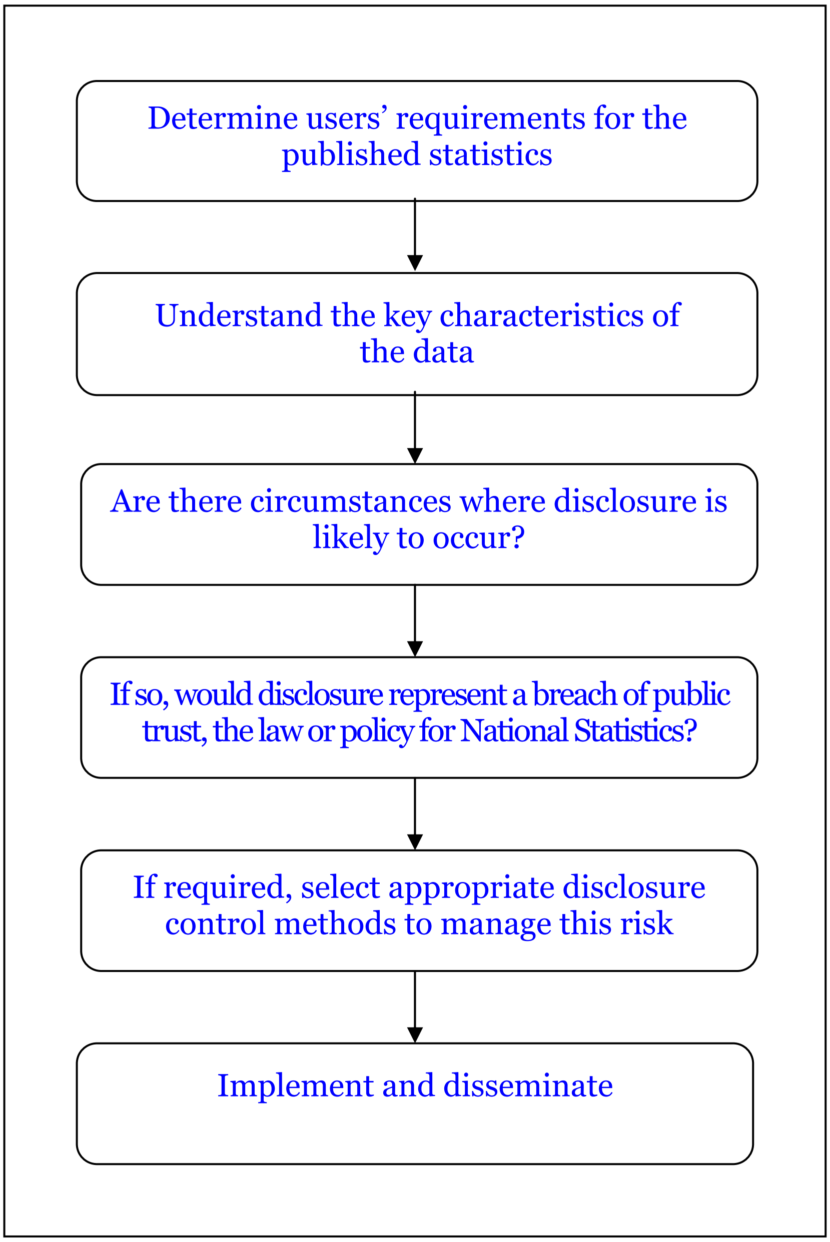
> Through trust in National Statistics, participants will see that data collection is necessary, that they do not risk being identified and that there is a clear benefit, to themselves and others, personally and as citizens, from the production of relevant and trustworthy statistical information. (National Statistics Code of Practice, Foreword)

Figure 1 shows the main steps to be taken in considering disclosure control in relation to tables of health data. It forms the structure for the guidance in this document.

- The first step involves establishing the user requirement for a particular health statistic.
- The second step involves gaining an understanding of the data that will underpin the statistics. The characteristics of the data will affect any disclosure risks. In particular, risk increases as statistics become more detailed (in terms of geography and categories) and as the dimensions of the table grow. Risks are higher if the distribution of the counts is skewed or the data are considered sensitive. Understanding the data will also involve establishing whether or not the data provider has the authority to disseminate the data.
- An assessment of disclosure risk should then be made. This will involve identifying situations where there is a likelihood of disclosure.
- Where a risk is identified, it is necessary to establish whether any disclosure would constitute a breach of public trust, of a legal obligation, or of a national or international policy standard for official statistics.
- If such a breach is thought to be likely, disclosure control methods can be used to manage the risk effectively. The various methods have different advantages and disadvantages and must be chosen bearing in mind users, uses and characteristics of the data.
- The final stage in the process is implementation of the methods and dissemination of the statistics.

For many published tables, the risk of identifying individuals will be minimal and no disclosure control methods necessary. Sometimes the information at risk of disclosure will not require protection for any reasons of public trust, the law, or National Statistics policy. For other information the issues may be more complex. No single solution is available for these instances. Instead, guidance is provided, based on the steps illustrated in figure 1, on how to develop solutions for different types of datasets. The guidance will allow data providers to develop their own confidentiality methods for different health statistics. These rules can then be applied to all published tables from a particular data source. This approach was adopted for the first stage of this review, in developing guidance on disclosure control for the publication of abortion statistics (ONS 2005b).

Figure 1: Main steps for ensuring access to non-disclosive statistics


Determine users' requirements for the published statistics
Understand the key characteristics of the data
Are there circumstances where disclosure is likely to occur?
If so, would disclosure represent a breach of public trust, the law or policy for National Statistics?
If required, select appropriate disclosure control methods to manage this risk
Implement and disseminate

4 Determining user requirements

> Users' views are essential in ensuring the relevance of National Statistics. (National Statistics Code of Practice, Principle)

Producers of statistics should design publications according to the needs of users, as a first priority. It is vital to identify the main users of the statistics, and understand why they need the figures and how they will use them in detail. This is necessary to ensure that the design of the output is relevant and the amount of disclosure protection used has the least possible adverse impact on the usefulness of the statistics.

The demands on health statistics are wide-ranging and include questions such as:

- what are the main health problems in the country or my town
- what is the quality of care in my local hospital and how does it compare with other hospitals
- how extensive are smoking, alcohol and drug-related problems, and which groups of the population are most affected, and
- are specific diseases becoming more or less common and are there groups of the population who are more affected than others?

Statistics help planners and managers of services understand the pressures on those services and how they are best organised for the benefit of users. They help decide how the substantial funds that go into health services are distributed and on what they should be spent. They also provide the basis for much research into health issues and treatments.

5 Understanding the key characteristics of the data and the required outputs

It is important to have a good understanding of the data that may require protection to assess any risk of disclosure. Here is a list of issues to take into account:

- The source of the data may affect the need to protect confidentiality. Health statistics are generally derived from registration processes, health-care sources, GP consultations, hospital data, waiting time records, etc
- Sensitive variables may require special attention
- The age of the data may reduce the risk of disclosure since the population of the statistic will change over time and become less identifiable. It is not possible to be more specific about this reduction in risk since it will differ between datasets and the populations represented
- The quality of data may determine the way in which the data are presented, the method of disclosure control or modify the need for disclosure protection
- Statistical units are defined as individuals, households, medical practitioners or health establishments. It is important to asses which units are represented in the data and are to be protected
- Particular issues may arise when the same unit is represented more than once. For example, if protection is required for practitioners, then cells in a table where all the values relate to a particular one, could be disclosive. Note, protection for practitioners is only required in certain circumstances, eg by Abortions Regulations. Further information concerning practitioner confidentiality may be obtained from the Office of the Information Commissioner at www.ico.gov.uk
- 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
- The disclosure risk for event-based data will be different from residence-based data. In order to identify an individual in a table for patients visiting a health clinic one would need to know that the individual is included in the population base for the table, ie has attended the clinic. The risk reduces if the population base or coverage of the table is not easily identifiable

It is also important to consider the characteristics of the tables. Where tables are very simple and presented at a high level of aggregation (including geography), disclosure issues are unlikely to arise. When tables become more detailed, and the counts in individual cells are small, the risk of identification may increase and protection may be needed. If the spread of values is skewed across a table, the risk in particular cells may increase above an acceptable level.

Issues may arise with linked tables where risk of disclosure can increase by differencing or through combining with other data. One particular problem that can occur with multiple or linked tables 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 of a potentially disclosive count. For health statistics this may occur when tables are produced from the same dataset for two non-coterminous geographies, eg Primary Care Organisation (PCO) and Local Authority District (LAD) in England. More details are provided in the working paper on risk assessment.

6 Assessment of disclosure risk for the intended statistical outputs

> For any National Statistic, statistical disclosure control measures will be adequate to ensure the confidentiality guarantee, and beyond that, as comprehensive as can be achieved without unduly compromising relevance, integrity and quality. (National Statistics Protocol on Data Access and Confidentiality)

In order to develop suitable confidentiality protection, a risk assessment should be undertaken. Risk is a function of likelihood (related to the design of the table), and impact of disclosure (related to the nature of the underlying data). Decisions on likelihood and impact should be made by those who have a detailed understanding of the statistics and experience of the interest in the figures. It is important to consider the views of patients and carers in the assessment of the impact of potential identification. In order to be explicit about the disclosure risks to be managed one should consider a range of potentially disclosive situations and take action to prevent them. The situations should be used to identify those parts of the statistical table that could lead to disclosure, termed ‘unsafe’ cells (commonly, cells containing small counts). Appropriate confidentiality rules should be applied to these cells. It is not possible to protect against all risks, this is a risk management not a risk elimination exercise. Three example situations are described in more detail.

General attribute disclosure

General attribute disclosure arises when someone who has some information about a statistical unit could, with the help of data from the table, discover details that were previously not known to them.

Example

A table of statistics for psychiatric services at a hospital shows admissions by single years of age, and diagnosis. Attribute disclosure has occurred if someone, who knows their neighbour was admitted for such service, discovers from the statistic that they are schizophrenic.

Disclosure may arise if there is a count of 1 in a marginal total (row or column) as in table 2, where a treatment of type 1, 2 and 3 is broken down by age bands. Anyone who knows that a particular individual under 12 has received a treatment would learn that it was a type 1 treatment. Attribute disclosure could occur from a count of 2 in a marginal total where one of the units may identify the other and thereby disclose further information.

Table 2: Treatment, by type and age

Treatment	Age				Total
	< 12	12–15	16–19	> 19	
Type 1	1	0	7	1	9
Type 2	0	0	18	19	37
Type 3	0	12	5	0	17
Total	1	12	30	20	63

Disclosure can also occur from cells with larger values, where they appear in a row or column dominated by zeros. A zero in population data allows one to say that no-one in the population has that attribute. This can be seen in table 2, which reveals that no 12–15 year olds are having treatment type 1 or 2 so all 12–15 year olds having the treatment are having type 3 treatment. The risk from many zeros within tables will not be significant, but, in some cases, they may need to be protected.

Disclosure risks may increase where groups of units that appear in the same table know enough about each other to identify each other and potentially discover something new. This can occur where units share characteristics or are grouped in some way, eg individuals from the same clinic or practitioners from the same hospital.

In order to protect against general attribute disclosure, at a minimum, care should be taken where rows or columns are dominated by zeros and in particular where a marginal total is a 1 or 2.

'The Motivated Intruder'

Data in a table is combined with information from local sources to identify a statistical unit and disclose further details.

Example

An intruder with a special interest in conception statistics discovers from a table that only a small number of very young women have conceived in a particular local area. The small number in the cell doesn't tell the intruder who the women are but it may prompt them to follow up other sources of information to locate the individuals and discover – and disclose - more details.

This situation may occur when small values are reported for particular cells. In a large population (for example, a country or region), the effort and expertise required to discover more details about the statistical unit may be deemed to be disproportionate. As the base population is decreased by moving to smaller geographies or sub-populations, it becomes easier to find units and discover information.

Although the local sources reveal the identity of the individual it is the statistics that cause the motivated intruder to start looking and attempting to reveal what is disclosive. The Protocol on Data Access and Confidentiality outlines that one does not need to take into account all local sources but information *likely* to be available to third parties.

In order to protect against a motivated intruder, at a minimum, all cell counts of 1 or 2 for geographies below LAD level or PCO in England and Local Health Board (LHB) in Wales or equivalent are defined as unsafe. As an indicator PCO sizes in England range from 64,200 to 369,800[1] and the average population size is 165,300. LADs in England range from 2,200 to 992,400[1] and the average population size is 141,100. This geographic level is provided as a general guideline to reflect that disclosure risk increases with smaller geographies. There may well be instances where some areas below this level are quite large and do not pose a particular risk under this scenario.

Identification and self-identification

Where a cell has a large value, risks arising from identification are not usually significant. Where a cell has a small value, particularly if the count is 1, this does need more consideration as identification or self-identification can lead to the discovery of rareness, or even uniqueness, in the population of the statistic. Hence there is a difference between being able to say that someone belongs to a population in a cell with a value of say, 162, and being able to say that a particular named person is the individual in a cell with a value of 1. For certain types of information, rareness or uniqueness may encourage others to seek out the individual. The threat or reality of this could cause harm or distress to the individual, or may lead them to claim that the statistics are inadequate to protect them, and therefore others.

Example

A statistic showing attendance at a drug misuse clinic by age and sex has a count of 1 for a particular ward. The individual may in fact be the only person who knows who this 1 is but they may feel exposed by the statistic. If this fear is communicated to their peers, it may spread, and the result may be a lack of trust in the confidentiality of their use of the clinic.

Identification or self-identification will occur from any cells with a count of 1, representing one statistical unit. The same is true of cells with a value of 2 representing two units, where one of the units contributing to the cell may identify the other. This could occur when groups of people or organisations who know enough to identify each other appear in the same table, eg individuals from the same clinic.

In order to protect against unique identification/self-identification, at a minimum all cells of size 1 or 2 are usually considered unsafe. Although direct identification/self-identification is not necessarily a significant

[1] Based on the mid-2004 population estimates

risk, protection is often required since identification can lead to attribute disclosure when more than one table is disseminated from a data source. The identified individual in an internal cell of a table can become a marginal cell in another table and a new attribute could be learned.

Risk categories

A risk assessment exercise should be undertaken to develop suitable confidentiality rules for different datasets. In practice it is likely that producers of statistics will find that outputs can be placed into one of three broad risk categories, defined in terms of the likelihood of an attempt to identify individuals, and the impact of any identification. Decisions on the likelihood and impact of identification should be made by those with a detailed knowledge of the data.

Medium Risk: In order to ensure protection from the disclosive situations described above for the majority of health statistics it will be sufficient to consider all cells of size 1 or 2 unsafe. Care should also be taken where a row or column is dominated by zeros.

High Risk: For some health statistics the likelihood of an identification attempt will be higher, and the impact of any successful identification would be great, eg statistics on abortions, AIDS/HIV, STDs. In order to ensure protection all cells of size 1 to 4 are considered unsafe and care should be taken where a row or column is dominated by zeros. Higher levels of protection may be required for small geographical levels or for particular variables with an extremely high level of interest and impact. Recommendations for abortions statistics have already been published (see www.statistics.gov.uk). Where the dissemination of certain data is covered by legislation, special care may need to be taken to ensure that the provisions of the legislation are met. It is likely that some of this data will fall into this category.

Low Risk: For some health statistics the likelihood of an attempt at identification may be considered to be lower if tables are disseminated at a high level of aggregation and only limited tables are produced from the one database, ie no risks from linking between current and future releases. A high level of aggregation reflects that the likelihood of disclosure decreases as the size of the population of the statistic increases. Also in the majority of cases tables produced below this level contain a high proportion of unsafe cells and the utility of such a table would be diminished. The level of aggregation will depend on the detail of the variables spanning the table. As a guide, in terms of geography a high level of aggregation should be interpreted as LAD or PCO in England and LHB in Wales (or equivalent) and above. Likelihood of an attempt at identification will also decrease if the population base or the coverage of the table is not easily identifiable. If the impact of any successful identification is also low then the heath statistics in this category will not usually require any protection beyond good table design. However, in order to prevent attribute disclosure care should be taken where rows or columns are dominated by zeros and in particular where a marginal total is a 1 or 2.

The likelihood of an attempt at identification and its impact may be heightened and additional protection required if:

- any other disclosive situations are likely to occur
- statistical units are represented more than once in the table. The likelihood of identification may increase since larger cells in the table may be associated with one statistical unit, eg if the statistical unit is a patient and the table reports annual hospital admissions, then a cell of 4 could represent the 4 times the patient was admitted
- groups of statistical units are represented in the table, eg patients from a particular clinic
- tables based on the dataset have already been released. The likelihood of identification may increase due to linking and differencing with these past releases. For large databases protecting against this risk may not be a trivial exercise, and
- other freely available datasets can be linked to the tables

These guidelines replace previous practices that have been adopted within the health field, such as the rule of thumb to suppress all values in tables less than 5. This guidance outlines the main steps to be taken in considering disclosure control allowing different solutions to be developed for different datasets taking into account detailed risk assessment and the latest disclosure control methods.

Some public authorities are required by law to provide partly or fully disaggregated data to the public either through statutory reports, or upon application. For example, the Registrar General is required by law to provide, upon reasonable request, certificate copies of death registrations which contain the name and recorded cause of death of the deceased. Provided published statistics do not allow for the discovery of other confidential information, it is generally acceptable for them to allow for the discovery of information equivalent to that otherwise required to be made publicly available in statutory reports or upon request. ONS has published its advice on vital statistics (ONS 2005c).

More details on the risk assessment process are provided in the second working paper.

7 Does the disclosure risk constitute a breach of statistical obligations?

> Where data are collected or used for statistical purposes, we guarantee to protect confidentiality. (National Statistics Code of Practice)

When establishing whether confidentiality protection is required for a particular health statistic, it is necessary to consider public trust and co-operation, and legal rights and obligations, as well as national and international standards for statistics. Thus there are acceptable disclosure risks and unacceptable disclosure risks.

The production and use of health statistics depends on the co-operation and trust of citizens. Such trust cannot be maintained unless the privacy of individuals' information is protected. Failure to respect privacy might result in harm or distress to a specific individual. Sensitive personal records, therefore, need to be strictly confidential. On the other hand, there is a legitimate public interest in having ready access to statistical information.

The legal framework covering the use of personal health information is complex. When such information is transformed into statistics, the legal framework is much simpler. The statistical information can be widely and freely used provided confidentiality protection has been applied such that it is no longer likely that the information can be related to specific identifiable individuals.

When the information in a health statistic does not relate to an identifiable individual (either on its own or in combination with other information likely to be available), there can be no breach of the duty of confidence owed or any conflict with data protection or human rights legislation.

National and international standards for official statistics

It is a United Nations fundamental principle of official statistics that the records of individuals, businesses or events used to produce official statistics are kept strictly confidential. The National Statistics Code of Practice and Protocol on Data Access and Confidentiality both conform to this principle and provide the ONS policy framework for official statistics. The Code of Practice guarantees confidentiality to those who provide private information for the production of National Statistics:

> Statement of Principle: Where data are collected or used for statistical purposes, we guarantee to protect confidentiality.

The Protocol on Data Access and Confidentiality, which underpins this statement of principle, states:

> The National Statistician will set standards for protecting confidentiality, including a guarantee that no statistics will be produced that are likely to identify an individual unless specifically agreed with them.

The Protocol also provides some guidance on how the standards for protecting confidentiality should be set:

> Statistical disclosure control methods may modify the data or the design of the statistics, or a combination of both. They will be judged sufficient when the guarantee of confidentiality can be maintained, taking account of information likely to be available to third parties, either from other sources or as previously released National Statistics outputs, against the following standard: It would take a disproportionate amount of time, effort and expertise for an intruder to identify a statistical unit to others, or to reveal information about that unit not already in the public domain.

The final guidance published following this consultation may be taken as the National Statistician's standard for protecting the confidentiality of health statistics. The working papers and the report represent the recommendations for designing statistics so as to protect confidentiality while maximising the information in the outputs.

More details on the legal and policy considerations when considering the confidentiality protection for published health statistics are provided in the first working paper.

8 Selecting disclosure control methods

> Statistical disclosure control methods may modify the data or the design of the statistic, or a combination of both. (Protocol on Data Access and Confidentiality)

The cells identified by the procedures in sections 6 and 7 as posing an unacceptable risk of disclosure are 'unsafe'. Where required, disclosure control methods can be used to reduce the risk by disguising the unsafe cells. The choice of method must balance uses to be made of the information, simplicity of approach and the implications of different approaches from a patient/carer perspective.

The methods are divided into three categories those that determine the design of the table, those that modify the values in the table and those that adjust the data before tables are designed. Descriptions of each method with advantages and disadvantages are provided below. In addition examples where each method has been implemented are outlined. Each example dataset (other than the 1991 Census) can be found on the ONS Neighbourhood Statistics website (www.neighbourhood.statistics.gov.uk).

Table redesign is recommended as a simple method that will minimise the number of unsafe cells and preserve original counts.

Table 3: Statistical disclosure control methods – design the table

Method	Description	Advantages	Disadvantages	Examples
Table redesign	Disguise unsafe cells by: - grouping categories within a table - aggregating to a higher level geography or for a larger population sub-group - aggregating tables across a number of years/months/ quarters	Original counts in the data are not damaged Easy to implement	Detail in the table will be reduced May be policy or practical reasons for requiring a particular table design	Teenage conception statistics are published for Local Authority or higher and the City of London is combined with Hackney

If unsafe cells remain in the output tabulation, further protection methods should be considered in order to disguise them.

Table 4: Statistical disclosure control methods – modify cell values

Method	Description	Advantages	Disadvantages	Examples
Cell suppression	Unsafe cells are not published. They are suppressed and replaced by a special character, such as '..' or 'X', to indicate a suppressed value. Such suppressions are called primary suppressions. To make sure that the primary suppressions cannot be derived by subtraction, it may be necessary to select additional cells for secondary suppression	Original counts in the data that are not suppressed are not adjusted	Most of the information about suppressed cells will be lost Secondary suppressions will hide information in safe cells Information loss will be high if more than a few suppressions are required In order to protect any disclosive zeros, these will need to be suppressed Does not protect against disclosure by differencing Complex to implement optimally if more than a few suppressions are required, and particularly complex for linked tables	Statistics on low birthweight babies are protected using suppression
Rounding	Rounding involves adjusting the values in all cells in a table to a specified base. This creates uncertainty about the real value for any cell while adding a small but acceptable amount of distortion to the data	Counts are provided for all cells Provides protection for zeros Protects against disclosure by differencing and across linked tables	Cannot be used to protect cells that are determined unsafe by a rule based on the number of statistical units contributing to a cell Random rounding requires auditing; controlled rounding requires specialist software	Statistics on claimants of disability living allowance, incapacity benefit and severe disablement allowance are protected by rounding to base 5
Barnardisation	A post-tabular method for frequency tables where internal cells of every table are adjusted by +1, 0 or -1, according to probabilities	Protects against disclosure by differencing	High level of adjustment may be required in order to disguise all unsafe cells Will distort distributions in the data	Implemented for the 1991 Census

If a data provider has access to the individual record level data then disclosure control methods can be implemented that adjust the data before tables are designed.

Table 5: Statistical disclosure control methods – adjust the data

Method	Description	Advantages	Disadvantages	Examples
Record swapping	Swap pairs of records within a micro-dataset that are partially matched to alter the geographic locations attached to the records but leave all other aspects unchanged	Protects against disclosure by differencing	High level of swapping may be required in order to disguise all unsafe cells Will distort distributions in the data. Method not transparent to users	Used in combination with small cell adjustment to protect the 2001 Census for England and Wales

Alternative methods for presenting data can be considered as an approach for 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. These could include presenting data graphically or providing commentaries or analytical outputs. More details and examples are provided in the working paper on risk management.

9 Implementing the guidance

The proposed guidance will allow data providers to set disclosure control rules and select appropriate disclosure control methods to protect different types of health statistics that are to be published. The most important consideration is maintaining confidentiality but these decisions will also accommodate the need for clear, consistent and practical solutions that can be implemented within a reasonable time and using available resources. The methods used will balance the loss of information against the likelihood of individuals' information being disclosed. Data providers should be open and transparent in this process and document their decisions and the whole risk assessment process so that these can be reviewed.

When looking at published health statistics, users should be aware that the dataset has been assessed for disclosure risk, and methods of protection may have been applied. For quality purposes, users of a dataset will be provided with an indication of the nature and extent of any modification due to the application of disclosure control methods. Any technique(s) used may be specified, but the level of detail made available should not be sufficient to allow the user to recover disclosive cell counts. Examples of such statements can be found in the metadata for datasets disseminated on the ONS Neighbourhood Statistics website (www.neighbourhood.statistics.gov.uk).

The final guidelines will help develop confidentiality solutions for different types of health statistics. Data providers may need to make judgements in a wider context than the specific statistics that they are producing at a particular time. They need to be aware of decisions made by others within their organisation, either in the past or for similar sectors. It is important that decisions are set within this context to ensure consistency and applicability within the strategic and policy context of the organisation. Decisions also need to be made in the context of wider information governance arrangements both in a organisation and in the health and social care sector more widely.

When data are shared with a second party for the purpose of publication (the assumption being made that this sharing complies with any legal or policy requirements), providers will try to make sure that the second party follows the general guidance and any specific confidentiality rules that have been developed. This will ensure consistency between published statistics derived from the same source.

Any change in disclosure control rules for a health statistic raises the issue of possible changes to past releases. In general, new disclosure control rules will be implemented for future releases but no changes will be made to past releases. However, where disclosure rules are altered to allow more data to be released, and where resources allow, a provider can consider re-releasing past datasets with more detail.

Individuals have a general right of access to information held by public authorities, through the Freedom of Information (FoI) Act. Confidentiality policy developed using this guidance can be used to help decide which exemptions in the Act are relevant, and which should be cited when withholding confidential statistical information. Whilst it is good practice to explain a general policy for the withholding of information this must be done in addition to, and not in place of, the exemptions in the FoI Act. FoI requests should always be considered on a case by case basis. There may be cases when decisions about a case are different to the general policy for the publication of statistics. This does not mean that the policy is wrong since it has been developed for use in a production process. Whist confidentiality must always be maintained, a decision made under FoI to provide information in a form different to the published outputs is compatible with this guidance. More details can be found in the working paper on the legal and policy considerations for confidentiality protection.

Statistical confidentiality is a public interest which will normally outweigh other relevant public interest in disclosing the underlying confidential records to the public. There will be rare occasions when the public interest in disclosing the records outweighs the public interest in confidentiality, for example a requirement to publish an occurance of an infectious disease. Such decisions will only be taken at the highest level and in consultation with the National Statistician. Usually it will be found that the records it is in the public interest to disclose are not statistics, but factual information and therefore subject to a different set of rules or guidance.

It should be noted that under the Freedom of Information Act 'statistical information' and 'factual information' are treated differently within the Section 35 exemption. Guidance on this exemption can be found on the Department of Constitutional Affairs website (www.dca.gov.uk). Most simply, factual information is the records of events or administrative actions, and statistical information is the outcome of a transformation, aggregation or analysis of such records performed using a repeatable methodology. Thus the records of a disease are factual information, and the aggregation and analysis of those records is statistical information.

10 Sources of more information

This guidance outlines the issues concerned with protecting the confidentiality of health statistics and describes an approach for ensuring that the public interest in the use of the figures is met while managing data disclosure risks. It also spells out the main steps that a data provider will consider in order to develop specific confidentiality rules for different types of health statistics.

Technical advice on the issues raised in this paper can be found in the attached working papers:

- Confidentiality protection – legal and policy considerations
- Risk assessment
- Risk management
- Glossary, and
- References and other guidance

A working example of the approach is provided for one extreme case of sensitive data in the guidelines for abortion statistics (ONS 2005b).

More advice on this document can be obtained from ONS; send emails to info@statistics.gsi.gov.uk, subject: Disclosure Review.

References

ONS (2004) National Statistics Code of Practice www.statistics.gov.uk/about/national_statistics/cop/about.asp

ONS (2004) Protocol on Data Access and Confidentiality www.statistics.gov.uk/about/national_statistics/cop/protocols_published.asp

ONS (2005a) *Data Sharing for a Statistical Purpose: A Practitioners' Guide to the Legal Framework* www.statistics.gov.uk/StatBase/Product.asp?vlnk=14201&Pos=&ColRank=1&Rank=272

ONS (2005b) *Disclosure Review for Health Statistics. First Report: Guidance for Abortion Statistics* www.statistics.gov.uk/downloads/theme_health/abortion_stag_final.pdf

ONS (2005c) ONS policy on protecting confidentiality within birth and death statistics www.statistics.gov.uk/statbase/Product.asp?vlnk=5768