Government Statistical Service

GSS/GSR Disclosure Control Guidance for Tables Produced from Administrative Sources

October 2014

Introduction

The last few years has seen continued interest from users and producers in ensuring official statistics reach their full utility whilst ensuring the risk of disclosure is minimised.

As a result of this continued interest the Statistical Suppliers and User Group, a forum for engagement between producers and users of official statistics, commissioned the GSS to update its existing disclosure control guidance published in 2007. This revised guidance on disclosure control for administrative sources has been updated to reflect the views expressed by a range of users and producers. The previous version of this guidance was approved by the GSS and GSR and hence applies to both social researchers and government statisticians.

Statistics based on administrative data sources (which here include Census data) support a wide range of users and uses, providing essential information for government, business, academia and the community. Many of these areas of work require detailed figures, which may raise issues about data confidentiality. Producers of such statistics must ensure that their statistics meet the needs of users by enabling relevant analysis to be carried out while at the same time protecting confidentiality.

This guidance describes the approach that data providers should follow when producing standard outputs and any ad-hoc requests based on a general framework for addressing the question of confidentiality protection. For Freedom of Information (FOI) requests this document can be used in conjunction with the exemptions in the Freedom of Information Act 2000. This is discussed further in Section 4.

The following elements of the framework are described in the guidance with more details provided in the Appendix:

- Determining user requirements
- Understanding the key characteristics of the data and outputs
- Assessing disclosure risk
- Legal and policy considerations
- Disclosure control methods
- Implementation

Relevant associated documents are

The *Code of Practice for Official Statistics* See https://gss.civilservice.gov.uk/wp-content/uploads/2012/12/Official-Statistics-Code-of-Practice.pdf (CoP) and specifically Principle 5: Confidentiality,

National Statistician's guidance: Confidentiality of Official Statistics See https://gss.civilservice.gov.uk/wp-content/uploads/2012/12/Confidentiality-of-Official-Statistics-National-Statisticians-Guidance.pdf (CoOS guidance) set out principles for how to protect personal data from being disclosed.

Another useful document is titled *Using Administrative Data: Good Practice Guide for Statisticians*
https://gss.civilservice.gov.uk/wp-content/uploads/2012/12/Interim-Admin-Data-guidance.pdf
This interim document provides further guidance for producers of official statistics about administrative data and is designed to ensure that any published tables based on administrative sources comply with the CoP and CoOS guidance.

The *ICO Anonymisation Code of Practice*
http://www.ico.gov.uk/for_organisations/data_protection/topic_guides/anonymisation.aspx was published in 2012 by the Information Commissioner's Office and this provides considerable support and guidance to data providers charged with generating and publishing tables (or microdata) from source microdata.

The CoP states ‘Ensure that arrangements for confidentiality are sufficient to protect the privacy of individual information, but not so restrictive as to limit unduly the practical utility of official statistics.’ This emphasises the dual purposes of statistical disclosure control. Detail relating to individual statistical units is to be protected but the released data must still be of high practical utility for users.

This document will be updated regularly to accommodate the latest thinking in statistical disclosure control.

Contents

1. Purpose

This is one of a pair of documents. This one concentrates on tables produced from administrative data sources while the second one (also available from the same source) gives guidance on tables from sample surveys.
Statistics based on administrative data sources (which here include Census data) support a wide range of users and uses, providing essential information for government, business, academia and the community. Many of these areas of work require detailed figures, which may raise issues about data confidentiality. Producers of such statistics must ensure that their statistics meet the needs of users by enabling relevant analysis to be carried out while at the same time protecting confidentiality.

This guidance describes the approach that data providers should follow when producing standard outputs and any ad-hoc requests based on a general framework for addressing the question of confidentiality protection. For Freedom of Information (FOI) requests this document can be used in conjunction with the exemptions in the FOI. This is discussed further in Section 4.
The following elements of the framework are described in the guidance with more details provided in the Appendix:

- Determining user requirements
- Understanding the key characteristics of the data and outputs
- Assessing disclosure risk
- Legal and policy considerations
- Disclosure control methods
- Implementation

The *Code of Practice for Official Statistics*[1] (CoP) and specifically Principle 5: Confidentiality, and the *National Statistician's guidance: Confidentiality of Official Statistics*[2] (CoOS guidance) set out principles for how to protect personal data from being disclosed. Another guidance document is titled "*Use of Administrative and Management Information*" and this provides guidance for producers of official statistics based on administrative data. This guidance is designed to ensure that any published tables based on administrative sources comply with the CoP and CoOS guidance. The *ICO Anonymisation Code of Practice*[3] was published in 2012 by the Information Commissioner's Office and this provides considerable support and guidance to data providers charged with generating and publishing tables (or microdata) from source microdata.

The Code of Practice for Official Statistics states 'Ensure that arrangements for confidentiality are sufficient to protect the privacy of individual information, but not so restrictive as to limit unduly the practical utility of official statistics.' This emphasises the dual purposes of statistical disclosure control. Detail relating to individual statistical units is to be protected but the released data must still be of high practical utility for users.

The previous version of this guidance was approved by the GSS and GSR and hence applies to both social researchers and government statisticians. This version has been further updated following comments from a range of interested parties.

[1] See http://www.statisticsauthority.gov.uk/assessment/code-of-practice/index.html
[2] See http://www.statisticsauthority.gov.uk/national-statistician/guidance/index.html
[3] See http://www.ico.gov.uk/for_organisations/data_protection/topic_guides/anonymisation.aspx

2. Scope

This guidance applies to frequency tables (tables of counts) derived from registration processes or administrative sources, which have a near-complete coverage of the population or a sub-population. It includes tables produced from UK Census data but not subsamples of Census data. The ONS has produced previous guidance for Health Statistics, Neighbourhood Statistics[4] and Birth and Death Statistics (revised in 2014)[5] while the Department of Health is updating the abortion statistics guidance. This guidance is a continuation of the earlier documentation but reflecting the current data transparency agenda.
Documentation is also available for specific outputs with one example being the Anonymisation Standard for Publishing Health and Social Care Data Specification which came into effect on 30th April 2013[6].
The aim is to ensure that a standard approach is used to avoid the publication of data which could identify individuals in NHS produced data. This Standard is specific to NHS data releases rather than general statistical publications as is the case with this GSS document.

Advice such as this specialist Health and Social Care document and the forthcoming NHS abortion statistics guidance ought to be the first point of reference for the relevant outputs with these guidelines to be used for other specific data requests on a case by case basis.

The NHS Anonymisation Standard is specific to releases of health and social care data, and goes wider than statistics. This GSS guidance is specific to the release of statistics but applies more widely than health and social care. For the overlap (statistics on health and social care) either may be used, provided a reasoned decision is made.

Guidance for tables produced from surveys is also available. Updated guidance for survey data is being released at the same time as this guidance.

Tables from administrative sources include those:

- released by the GSS/GSR to the public through normal publications and customer requests. This includes tables sourced outside GSS/GSR, but published in GSS/GSR reports
- released through the Freedom of Information Act (FOI) (2000) and in Scotland the Freedom of Information (Scotland) Act or FOISA (2002). This document 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 (FOISA) Act.
- derived from microdata or other non-publishable data accessed by licensed users.

Exemptions and exception arrangements for this policy are detailed in Section 5.

[4] http://www.neighbourhood.statistics.gov.uk/HTMLDocs/images/NeSS_data_access_tcm97-51092.pdf
[5] http://www.ons.gov.uk/ons/guide-method/best-practice/disclosure-control-policy-for-birth-and-death-statistics/index.html
[6] See the documents in this link http://www.isb.nhs.uk/documents/isb-1523/amd-20-2010/index.html

3. Key Steps

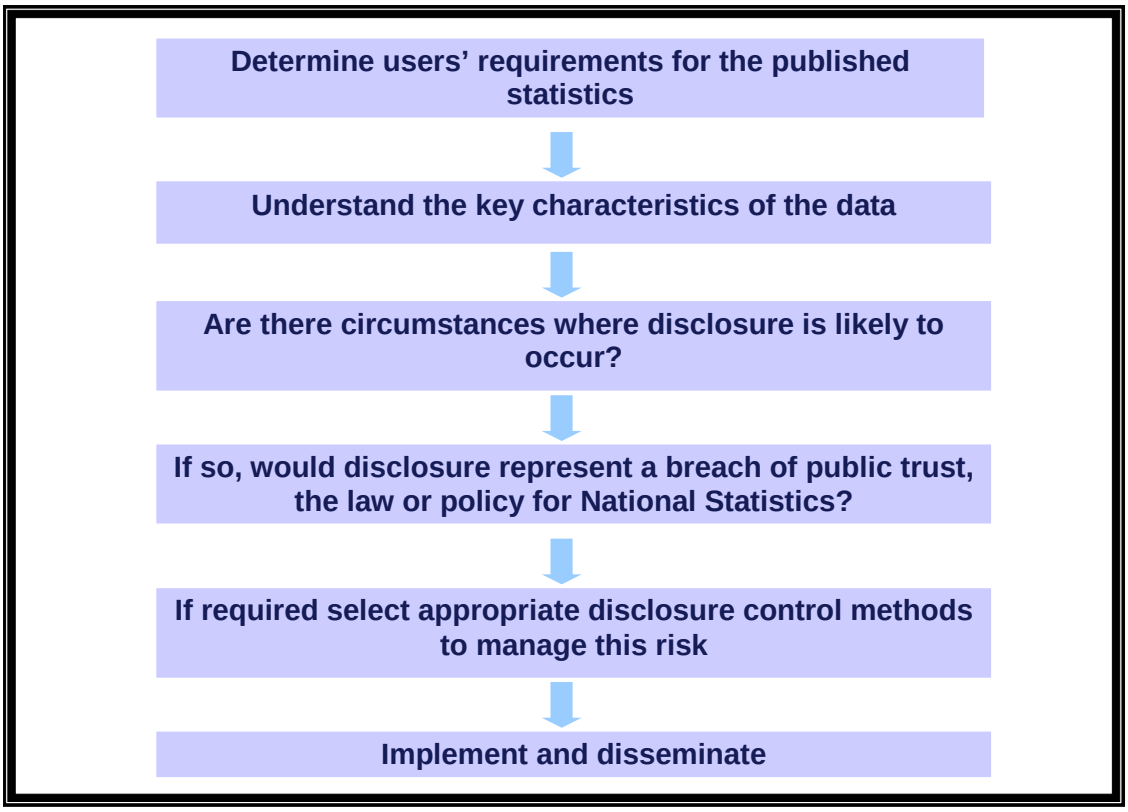

Figure 1: Main steps for ensuring access to non-disclosive statistics

Figure 1 shows the main steps to be taken in considering disclosure control in relation to tables of administrative data. These steps describe the process to be followed in general terms. The steps are broadly defined so as to reflect the prevailing pressures for publishing data. For example, the current era of open data can be reflected by taking a more rigorous approach to step 3 above.

- The first step involves establishing the user requirement for a particular statistic. In particular identify the main users, and find out why they need the statistics and how they will be used. Under open data conditions knowing this detail may not be necessary as fewer restrictions are placed on use of the data.
- 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.
- 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 the choice must bear in mind users, uses and characteristics of the data.
- The final stage in the process prior to publication is implementation of the methods and dissemination of the statistics.

For many tables, 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. There is no one solution 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. This guidance will allow data providers to develop their own confidentiality methods for different statistics. These methods can then be applied to all published tables from a particular data source. To ensure confidentiality of outputs, the selected method must be applied to all tables created from the same underlying microdata.

4. Guidance

The reason for releasing administrative tabular data is to enable researchers and the general public to use the outputs for their own purposes. Releasing any data into the public domain carries an element of risk that disclosure could occur, but this is no reason not to publish. If data utility can be kept as high as possible while a low level of risk is maintained and the reasons for this can be justified both producers and users of the data should be content. Any data published should also be for the public good.

Sections 1 and 2 of this document show the range of internal guidance and Parliamentary Acts relating to the release of information (particularly personal information) into the public domain. All guidance documentation needs to be read in conjunction with the relevant Acts. Some background on the Acts is useful to see how they relate to the disclosure control process.
The legal underpinning to the release of data is the Data Protection Act (DPA) (1998)[7]. This outlines the rights of the individual with respect to their personal data. The Act describes how these data can be used and maintained by public bodies. Particular variables are defined by the Part 1 section 2 of the Act as sensitive personal data.
These are:

- Racial or ethnic origin
- Political opinions
- Religious or other similar beliefs
- Trade Union membership
- Physical or mental health or condition
- Sexual life
- Commission or alleged commission of an offence
- Proceedings for any offence committed or alleged to have been committed

As can be seen sensitive personal data as defined by the DPA is wide ranging. Tables containing details of personal health and possibly tables relating to financial affairs would be defined as sensitive data.

There is little detail in much official guidance as to how personal information is classified. The variables above are useful attributes to consider when producing tables.

The FOI Act[8] also can be invoked when specific data are requested. Current Statistics Authority policy is for outputs from all FOI requests to be made public, therefore the

[7] http://www.legislation.gov.uk/ukpga/1998/29/contents
[8] http://www.legislation.gov.uk/ukpga/2000/36/contents

approaches to be followed when confidentialising outputs ought to e the same as for general requests. There are a number of exemptions where FOI requests can be refused. One of these is where the data are personal information. Part II of the Act gives background on the exemptions and Section 40 provides details on the personal information exemption.

Common law allows individuals to bring legal proceedings on the basis that the collection, use and disclosure of personal information is in breach of the obligation of confidence. Essentially the common law provides that anyone who receives confidential information must not disclose it without consent or justification. The common law duty of confidentiality extends to confidential information about deceased persons.

Ultimately the Information Commissioner is responsible for ensuring that good practice with respect to publishing data is followed. He/she is likely to have a view on any refusals to publish on confidentiality grounds.

Alongside the legal considerations the practical day to day aspects of disclosure control also need to be considered.
The variables defined in the DPA as being sensitive will be of high impact if disclosed. In general disclosure risk is a constant problem when personal data are released. Risk can be defined as the likelihood of an individual, business or household or other statistical unit (or related attributes) being identified in a published table. If the risk involves particularly sensitive variables that could cause excessive and unnecessary harm or distress to particular individuals or groups then the impact of any disclosure is increased.
Note that the term ‘impact’ refers to ‘statistical impact’ (the impact that disclosure of information will have on the individual, business or household concerned) in order to distinguish it from ‘reputational impact’ which occurs when a dataset is lost in transit or data are released inadvertently.

Levels of risk can be defined for outputs from administrative data. Examples of medium and high risk are shown below. The guidance here is an indication of how low cell counts can be disclosive especially at different levels of geography. The effect caused by variable sensitivity is discussed in Appendix A3, in particular Section A3.4. The values selected in the risk categories may appear to be arbitrary. However the key point to think about here is how an attacker would approach a table containing low frequencies.
If they consider the variables to be of great personal or political interest they will invest more time in attempting to find an individual or an attribute relating to an individual in the table.

This leads to a distinction being made regarding the risk level of the table. It may be considered to be an ill defined distinction but expert knowledge of the data should help to provide an informed outcome.
Low counts in the margins of a table are a particular issue not only because the row or column contributing to the margin will consist of small counts but also because the marginal total will highlight the small number of contributions to one particular variable and give encouragement to an attacker to investigate further.

For frequency tables, cells with low counts are likely to be the major problem. Recent documentation (such as the Anonymisation Standard for Publishing Health and Social Care Data Specification previously mentioned) use the term k-anonymity. This term can be used to refer to tabular or microdata and is a criterion for ensuring that there are at least k records in the data that have the same quasi identifier values. For example if the quasi identifiers were. Age Group, Gender and Ethnicity and k equalled 3, each combination of the variables would

consist of at least 3 records.). Any table in the cell with a count of 1 or 2 would therefore be disclosive.

Case studies in Appendix A3 show practical applications of these risk scenarios.
The release of sensitive personal data as defined by the DPA will have a greater if disclosed. This needs to be taken into account when considering disclosure risk. If a variable is especially sensitive or concerns very personal information such as the detailed medical history of an individual, any resulting identification could have a great impact on the individual being identified.

1. Low Risk: For some administrative statistics the likelihood of an attempt at identification may be considered to be low if tables are disseminated at a high level of aggregation such as National or Regional level and only limited tables are produced from the one database, i.e. no risks from linking between current and future releases. A high level of aggregation reflects a reduction in disclosure risk as the size of the population of the statistic increases. Statistics in this category will not usually require any protection beyond good table design.
However, care should be taken where rows or columns are dominated by zeros and in particular where a marginal total is 1 or 2.

2. Medium Risk / High Risk Many administrative statistics disseminated at a lower levels of aggregation, i.e. small geographies or small populations, or where many linked tables are produced from the same data set it is likely that 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.
There is similar recommendation for high risk outputs where identification attempts may occur frequently. Here the impact of any successful identification would be great, e.g. statistics on abortions. In order to ensure protection all cells of size 1 to 2 are considered unsafe and care should be taken where a row or column is dominated by zeros. High risk tables should also be looked at closely if they are especially sparse, i.e. a low cell average and containing many zeros. It may be that higher levels of protection may be required for small geographical levels or for particular variables with an extremely high level of interest and impact.

Zeros
These risk levels refer to the importance of zeros. Disclosure risk will be greater if zeros are distributed in particular ways. If all cells in a row or column are zero apart from one, an attacker would know that all members of the row belong to a particular category for the column variable. This is a form of group disclosure.
The distinction between structural and non structural zeros is also important. Structural zeros are those where the counts cannot be anything other than zero, such as 8 year old mothers whereas non structural zeros occur because nobody with that combination of characteristics is present in the population (although there is no practical reason why not). Risk in tables with zeros is generally determined by cells which contain non structural zeros.

These risk levels are discussed as part of the Case Studies in Appendix A3. where in some situations the recommended levels of protection may need to be increased particularly in relation to population at risk and the sensitivity of the variables.
Population at risk refers to the underlying number of people eligible for presence in a particular cell in a table (i.e. those sharing the characteristics of those in the table. The likelihood of disclosure will increase as the population at risk decreases. For a smaller population at risk there is greater possibility of disclosure). For example if a cell represented the number of women aged 18-34 who had given birth in a Local Authority, the population at risk would be all women aged 18-34 in that Local Authority. Practical examples of how the

overall table risk is related to the population at risk and variable sensitivity are shown in Appendix A3.4.

The NHS Anonymisation Standard for publishing Health and Social Care data describes different levels of population at risk to those in this document, for specific case studies. The guidance in this document is applicable more widely than health and social care while the NHS standard is specific to health and social care data but covers a wider area than just statistics. When this document and the NHS Standard overlap either may be used, provided a reasoned decision is made.

A table is 'unsafe' if it contains one or more cell with an unacceptable risk of disclosure . Disclosure control methods should be used to reduce the risk by modifying these cells. The guidance in Appendix A provides description of each method with advantages and disadvantages and examples. Table redesign is recommended as the initial method of disclosure control but should be balanced against user needs and publication plans. If further disclosure control is required then either controlled rounding (if suitable software is available) or cell suppression are the suggested options although other disclosure control methods are available. Careful judgement will be required when applying any method in order to ensure that the data are not damaged too much.

The aim for the vast majority of outputs is to introduce sufficient uncertainty into the data giving an element of doubt to any disclosure. This is an alternative to the case of zero risk where the data are protected so that disclosure is not just highly unlikely but impossible. This may be a requirement for some outputs but, in order to achieve this, considerable damage will need to be applied to the data this reducing data utility considerably.

Uncertainty relates to the extent to which cell values and apparent attribute disclosures in tables do not represent real respondents or real attribute disclosures. Published cell counts may not represent the true counts for a variety of reasons, some of which cannot be quantified easily (e.g. respondent or data capture error, mis-classification), and some that can be quantified (e.g. non-response and edit imputation, pre- or post-tabulation as part of disclosure control). The level of uncertainty deemed 'sufficient' is a matter of judgement and should be a general policy decision made by the data owner, who might take advice from their Head of Profession (where possible) and/or from ONS SDC Methodology Branch. Factors to take into account in setting the level include the amount of sensitivity in the data and the likelihood and impact of a real disclosure claim. Case studies and examples of the relative impact of particular variables are shown in the appendix. This should assist the data owner in making a decision

5. Implementation and Evaluation

In accordance with CoP Principle 5, Practice 4, and as outlined in the CoOS guidance paragraphs 36-40, implementation will achieve the obligation to protect against disclosure but will, through choice of disclosure method, strive to give the greatest practical utility possible in the released statistics.

Standard wording should be used in dissemination releases. 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 should be provided with an indication of the nature and extent of any modification due to the application of disclosure control methods but the level of detail made available should not be sufficient to allow the user to recover disclosive cell

counts. An example of standard wording can be in seen in the Conception statistics for England and Wales 2011. The metadata state:

Data displaying conception statistics as counts and rates must not be used to disclose information on abortions. Therefore, for conceptions leading to abortions, counts less than 10 and rates based on fewer than 10 events have been suppressed. To protect confidentiality for conceptions data, all counts lower than 5, and all rates based on fewer than 5 events have also been suppressed.

Occasionally it has been necessary to apply a secondary suppression to avoid the possibility of disclosure by differencing.

- 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 internally. Such a document should be structured using the key principles set out in Section 3.Public documents on the disclosure control method(s) applied should include details on the methodology while not disclosing specific parameter values.

- When releasing data the latest disclosure control rules should be applied. However changes to these rules do not necessarily imply that corresponding changes need to 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 as long as this does not compromise any outputs by differencing.

Situations where the GSS/GSR guidance may not apply, and any approval processes required, are given below.

Exemptions. Exemptions required by legislation from application of disclosure control must be listed on the Register of Exemptions, see Principle 5, practice 5, and the CoOS guidance, paragraphs 41-45 and Annex C. In addition to documenting all such cases, there will be circumstances where authorisation may be required from the National Statistician or Head of Profession.

Reduction in Confidentiality Protection. If none of the exemptions in Annex C of the CoOS guidance applies but a data provider wishes to provide less confidentiality protection than that described in this guidance, approval is normally required from the Head of Profession. Further advice can also be sought from the ONS Data Stewardship Group (DSG), the Legal Services branch (part of the Organisational and Capabilities Directorate) and the Statistical Disclosure Control branch (part of the ONS Strategy and Standards Directorate),

Access to disclosive tabular data. Principle 5, practice 2 and paragraphs 23–29 of the CoOS guidance cover the principles and arrangements for access to all disclosive (i.e. identifying) data, including disclosive tabular data as well as microdata. At the ONS the release of disclosive tabular data and microdata must be approved by the Microdata Release Panel. Each Government Department will have their own procedures when allowing access to disclosive tabular data and microdata. In the devolved administrations approval will be required from the Chief Statistician.

Mixture of pre and post tabular processes. Particular methods of applying statistical disclosure control are applicable to both tables and microdata. A number of Government Departments and other data suppliers allow public access to an interactive tabulation tool

which can be used to generate user defined tables. It is likely that this form of dissemination will become more popular in the future allowing specific detailed tables to be extracted. Techniques have to be applied to ensure that these tables are not disclosive. Generally the microdata are protected to some degree before the tables are generated. This approach will be discussed in the revised microdata guidance due for release in 2014.

6. Responsibilities

Each publication has an associated Responsible Statistician who is responsible for confidentiality protection of released data, ensuring that the standard disclosure control methods are applied, and any other special circumstances are taken into account. The disclosure control method will always be signed off by, or in the name of, the Head of Profession (not always the same personas the Responsible Statistician) which for certain releases may be the National Statistician or the Chief Statistician in a devolved administration.
Day to day management of disclosure control for data release may be delegated to output managers, data managers or others responsible for the confidentiality guarantee pertaining in outputs from administrative sources, whether the data are released by GSS or by others using data from this source.

ONS SDC Methodology team are happy to help and offer advice to users and data providers where necessary. Their contact email is
Sdc.queries@ons.gsi.gov.uk

Appendix A: General Guidance for Disclosure Control for Tables Produced from Administrative Sources

A1. Determining user requirements

The GSS/GSR produces a wide range of statistics based on administrative sources; examples are tables of statistics concerning crime, education, benefits and health.

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.

A2. 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.
- 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. Data quality can vary considerably due to many reasons. Possibilities include poor population coverage, information being poorly recorded during the processing of the data, a large proportion of imputed data being required and sensitive information not being given accurately by respondents.
- Typically statistical units are defined as individuals, households or businesses. It is important to assess 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 in a table or a series of tables.
- Disclosure risks may also increase if groups of statistical units (e.g. individuals from the same household) are represented in a table and, therefore, could potentially identify each other.
- The disclosure risk for event-based data will be different than residence-based data. For example 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, i.e. 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 less likely to arise. However, even at a high level of aggregation small cell counts can be a risk and the underlying population or sub population should be taken into account. 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 the 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 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 statistics this may occur when tables are produced from the same dataset for two non-coterminous geographies, e.g. wards and super output areas (SOAs).

A3. Circumstances where disclosure is likely to occur and managing this risk

In order to develop suitable confidentiality protection a risk assessment should be undertaken. This assessment should include factors such as the nature of the variables (risk and impact if released and the structure of the table (the number of observations in the table and their distribution).
Risk is a function of likelihood (related to the design of the table), whereas impact is related directly to the nature of the underlying data). Disclosure can be quantified in terms of both risk and impact. Disclosure risk is high when a table is designed so that there are cells in the table with low frequencies or when there are rows or columns where all the counts are in a small number of cells. Tables such as these could lead to identification of an individual and maybe disclose further information about this individual.
The impact of any disclosure is higher if the data are sensitive and great distress may be caused by releasing the data. As stated in Section 4 there is a list of variables defined as personal sensitive data by the Data Protection Act (1998). For example if a table of high risk was released giving details of treatment for mental health issues in a small geographical area it would have greater impact than a table of equal risk detailing use of local shops.

Decisions on risk and impact should be made by those who have a detailed understanding of the statistics and experience of the interest in the figures. 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 risk assessment should be reviewed on a regular basis as the tolerable level of risk may change over time. 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
- The motivated intruder
- Identification and self-identification

The related Case Studies discuss the practical decisions likely to be faced when producing tables from administrative data. There are no formal rules to follow when publishing tables but if the framework is followed secure outputs of high utility should be produced.

A3.1 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.

Attribute disclosure includes inferential disclosure, where information about a statistical unit can be inferred with a high degree of confidence.

Case Study 1

Please see the associated document ‘Case Studies for tables produced from administrative data’

A3.2 'The Motivated Intruder’

Data in a table are combined with information from local sources to identify a statistical unit and disclose further details. The level of disclosure risk that is tolerable may depend heavily on the sensitivity of the data.

Case Study 2

Please see the associated document ‘Case Studies for tables produced from administrative data’

The situation described in Case Study 2 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. Also the intruder is likely to have greater confidence in any claim s/he might make.

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 CoP Principle 5 practice 1 states that official statistics should not reveal the identity of any respondents with the risk of disclosure to include taking into account other relevant sources of information. These sources may be private or public but the relevance of them is determined by whether they are likely reasonably to be used to identify an individual and reveal information about them. Thus 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 Local Authority District (LAD) level or Clinical Commissioning Group (CCG) in England ought to be considered potentially disclosive. As an indicator CCG sizes in England range from 63,100 (NHS Corby) to 869,400 (NHS North, East, West Devon) (based on the mid-2012 population estimates). LADs (excluding the Isles of Scilly and City of London) in England range in population size from 37,369 to 1,073,045 (2011 Census outputs). The threshold value of 3 is chosen here to ensure that in a well defined geography (which would be of use to an attacker) there is neither an isolated individual nor a case where a member of a cell of frequency 2 may attempt to identify the other person in the cell. Larger cell values are likely to discourage the attacker.

Scotland, Wales and Northern Ireland have 14 Health Boards, 7 Health Boards and 1 Health and Social Care Board respectively. The populations here will be smaller than for English CCGs. All cells counts of 1 or 2 for lower geographies than these are potentially unsafe.

These levels of geography are 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. This is expanded upon in the discussion on 'Risk Category' (Section A3.4).

A3.3 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.

Case Study 3

Please see the associated document 'Case Studies for tables produced from administrative data'

Identification or self-identification will potentially 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 with similar characteristics who know enough to identify each other appear in the same table, e.g. individuals from the same household.

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

A further scenario which ought to be considered is that of the mosaic effect (or jigsaw effect) where different data releases, each of which may be deemed safe in isolation, can be pieced together to effect disclosure. The effect of many slightly different data releases is likely to become increasingly common with the ideas of Open Data and Transparency being widely promoted by the Government and more tables are requested through Freedom of Information. A hypothetical example of the mosaic effect is shown here.
Take an example of an administrative data set of 12 variables, four sensitive variables (S1-S4 of high impact), four visible variables (V5-V8) and four other variables (N9-N12), from which various tables are formed and released.
The sensitive variables could be
S1: Limiting long term illness (y/n)
S2: Sexual Identity
S3: Income
S4: Religion

Visible variables are those from which an individual can be categorised by observation
V5: Gender
V6: Ethnicity (Possibly difficult to identify correctly as the attacker's estimate of ethnic group may differ from how an individual categorises themselves)
V7: Household composition
V8: Occupation

Other variables
N9: Country of Birth
N10: Age in years
N11: Number of children
N12: Level of education

The dataset as a whole may be assessed as personal information in that it allows a person to be identified and other information, perhaps sensitive, to be revealed. The risk is that an intruder could use V5-V8 to identify an individual and find out something sensitive (S1-S4). The other variables N9-N12 can help as additional matching variables or catalysts.

Release 1: V5 V6 N9 N10 N11 N12 may be safe - Gender and Ethnicity may not be sufficient to identify a person

Release 2: S1 V7 N9 N10 may be safe - Household composition not sufficient

Release 3: S2 V8 N11 N12 may be safe - Occupation not sufficient

Release 4: S1 S2 S3 S4 N9 N12 may have sensitive information but unable to identify an individual (no visible variables)

It can be seen how these can build up. Country of Birth and Level of Education are common to Releases 1 and 4, Limiting long term illness and Country of Birth common to Releases 2 and 4, Sexual Identity and Level of Education common to Releases 3 and 4, Country of Birth and Education to 1 and 2. It is not straightforward for an intruder unless combinations of these variables approach uniqueness, but one can see the potential for approaching identification of an individual. It is worth remembering that linking and disclosure may only be possible for one or a small number of individuals, but that may be the individual with the most unusual combination of characteristics, and possibly, therefore, the most sensitive.

This mosaic effect requires considerable thought and each combination of outputs needs to be considered as a whole. The ultimate aim is to be pragmatic and not to think the worst case will always happen. As always attempt to balance the risk of releasing a group of linked tables with the utility these tables can provide.
Some steps to follow are:

Define the variables carefully. How sensitive are particular variables? What would be the impact of disclosure?
Are the tables released together or at different times? Are they ad-hoc requests to the same individual? This is where a log of releases would be of use. Releases to the same source would be likely to be more of a problem.
If a particular set of tables has been released previously it will be more difficult not to release in the future.

If a large number of linked tables are to be released from a particular dataset it may be advisable to protect the underlying microdata prior to tabulation. This is out of the scope of this document but protection of microdata from social surveys is being rewritten and will be published in 2014.

A3.4 Population at Risk Assessments

The effect of geography on disclosure risk was raised as part of the guidance in Section 4. This can be expanded to consider risk for different population sizes alongside variables of differing sensitivity.
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 output risk levels can be defined by the population at risk with a further breakdown by 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. When deciding on the particular category thought will need to be given both to the likelihood of an identification (based on population at risk as discussed in Section 4) and the impact that this identification would have. Tables with sensitive defining variables will almost certainly have a greater impact if an individual is identified correctly.

Impact through the release of specific combinations of variables ('statistical impact' as described in Section 4) is a subjective term but possible categorisations are shown below for a small number of variables commonly found in ONS and GSS outputs. Some of these are the same as the broader selection of variables defined as sensitive by the DPA.
As described in Section 4, variables of higher impact are more sensitive, meaning that any disclosure would cause a great deal of distress to the individual concerned.

High Impact Variables
Income
Racial / ethnic origin
Religious beliefs
Physical / mental health
Some types of crime (as a victim)
Sexual identity
Economic Activity
Industrial Classification (at the most detailed level)
Qualifications (level of qualification and details of subject)

The lower the level of geography at which a table is released the greater the level of risk for these variables. At Output Area level it is possible that variable combinations including one or more of the above variables will be disclosive.

Medium/ Low Impact Variables
Marital status (where not covered by vital registration)
Individual Age
Household composition
Geography at Local Authority or equivalent
Industrial Classification (grouped into a small number of categories)
Gender
Age group
House type

The risk will be lower if the geography at Region or equivalent

As stated previously, counts of 1 or 2 for tables not at a high level of geography need to be thought of as possibly disclosive. However as the populations of the PCOs and Health Boards differ widely, the population at risk for a table produced from a lower geography will not be consistent. It is difficult to give exact guidance but examples of population sizes are shown for the risk levels below. The relationship between risk and impact is also considered for 2 categories of impact, high and medium/low.

Population at risk ≥ 50,000

This is equivalent to the population of a small Local Authority or Unitary Authority with only a handful being smaller (e.g. Rutland with a 2012 mid year population of 37,015). Based on the 2012 mid year population figures Eden has a population of 52,651 and West Devon 53,859. The term Local Authority also covers Unitary Authority for the remainder of this section.
In a population of this size or larger any attempt at identification by an attacker would almost certainly not be successful. To relate an individual the table and relate it correctly to an individual in the population would be highly improbable.
Tables of low counts (1s and 2s in the margins) where one or more variable are high impact could be examined for disclosure issues but in almost all cases tables bases on this population at risk can be released with no application of disclosure control.

25,000 ≤ Population at risk < 50,000

This range is smaller than all but two Local Authorities[9] in England and Wales but higher than other defined geographies. However it can be thought of in terms of health statistics as being the female population aged 11-49 in a medium sized Local Authority. 2012 mid year populations estimates are 27,171 in Worcester, 36,461 in Middlesbrough and 42,981 in Chelmsford. .
An attacker attempting to determine that an individual in a table (for example of abortion statistics) was a relation or colleague would have to know a great deal about them to make a positive identification with great confidence.
If there is a high impact variable in the table then consider protecting counts of 1 or 2,although in many cases there will be no problems, otherwise publish with no additional protection.

12,500 ≤ Population at risk < 25,000

As with the previous risk range this can be thought of as being the female population aged 11-49 but here for a smaller Local Authority. Relevant populations are 13,232 in Ribble Valley, 16,324 in Copeland and 23,304 in Hartlepool. A similar methodological approach can be followed.
As the population is smaller tables with high impact variables will possibly require disclosure control to be applied for tables with counts of 1 or 2 although each case should be judged independently.

5,000 ≤ Population at risk < 12,500

The smallest Medium level Super Output Area (MSOA) population size is 5,000 with the average size being 7,500. (Examples are 7,608 in Bury007 and 8,126 in Birmingham128.) For a concentrated population of this size in an urban area an attacker could possibly identify a friend or relation in a table although some work would be required to do this with great confidence.

[9] Isles of Scilly and City of London have very small populations and are usually grouped together with other larger neighbouring Authorities

There is some risk for tables with at least one high impact variable so disclosure control may be appropriate for cells of size 1 or 2 in the body of the table and the margins and rows or columns dominated by zeros ought to be checked. Where no variables are high impact disclosure control may not be necessary although low counts especially in the margins should be looked at closely.

1,000 ≤ Population at risk < 5,000

Lower level Super Output Area (LSOA) min population sizes are around 1,000 with an average of 1,500. Examples are Birmingham 012E with a population of 1,635 and Greenwich003B where the population equals 1,739. At this population size discovering a friend or neighbour in a table becomes easier for an attacker. With a limited number of individuals who could be in the table it might only take a nosy neighbour a short amount of time to make a correct identification with great confidence.
If there is a least one high impact variable in the table disclosure control will be required for tables with cells of size 1 or 2 in the body of the table or the margins. Rows and columns dominated by zeros will also require the application of disclosure control. If there are no high impact variables low cell counts may require protection but each table will need to be looked at individually.

Population < 1,000

Tables based on populations of this size will be unusual. Maybe tables for a specific age group in a small geographic area will belong in this category. An attacker may be successful in finding individuals in tables based on a population this low.
Disclosure control should be applied to these tables irrespective of whether the variables are of high or low impact. Any cells of size 1 or 2 and row and columns dominated by zeros will require protection.

The levels of geography at which data are likely to be released (such as LAD) vary considerably in population size. This is why population at risk is used alongside the impact of a variable to define the risk of a table.
As with the cell values selected for risk levels in Section 4, the values for population at risk suggest levels at which an intruder might be more inclined to try to identify an individual in a published table. Smaller populations at risk require more protection. However these values are for guidance only (the document is for guidance and not a Standard which is expected to be followed).
Experts in particular areas of Official Statistics may want to choose different values for population at risk although maintaining the structure described above.

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, e.g. 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 same patient was admitted;
- groups of statistical units are represented in the table, e.g. individuals from a particular household;
- 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;
- other freely available datasets can be linked to the tables.

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 in law to provide, upon reasonable request, certificate copies of death registrations which contain the name and recorded cause of death of the deceased. (Note that the final cause of death is not public information). Providing 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. As previously stated ONS has published its advice on Birth and Death Statistics. Please note this advice has been updated during 2013/14.

A4. Does the disclosure risk identified constitute a breach of statistical obligations?

When establishing whether confidentiality protection is required for a particular statistic, it is necessary to consider public trust and cooperation, 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 statistics depends on the cooperation and trust of citizens. Such trust cannot be maintained unless the privacy of individuals' information and even a perceived risk is protected or shown not to be a true risk. 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 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 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 Code of Practice and CoOS guidance both conform to this principle and provide the GSS policy framework for official statistics. The CoP guarantees confidentiality to those who provide private information for the production of Official Statistics:

> Statement of Principle 5: Private information about individual person (including bodies corporate) compiled in the production of official statistics is confidential, and should be used for statistical purposes only.

The National Statistician's Guidance: Confidentiality of Official Statistics (CoOS), which underpins this statement of principle, states in paragraphs 7 and 8:

The code puts 'private information' into the scope of the confidentiality guarantee, where private information is that which

- relates to an identifiable legal or natural person and
- is not in the public domain or common knowledge, and
- if disclosed would cause them damage, harm or distress

Information available in the public domain does not become confidential information automatically when presented in a table or as another statistic. 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.

Other than to distinguish one unit from another for statistical purposes (for example, for data matching or linking exercises for statistical purposes, where identified data are essential for quality reasons), the statistician or researcher should have no interest in the individual statistical unit. In contrast, an intruder is someone who, for whatever reason, wishes to distinguish one statistical unit, in order to treat that unit separately and/or differently from the other statistical units in the dataset, for a non-statistical purpose.

The CoOS guidance paragraph 18 states that consideration of intruder scenarios "should be informed by the means likely reasonably to be used to identify an individual" and scenarios to be considered "will vary according to the topic of the statistic, its uses and other factors".

The term 'identify' is used frequently in legislation. For example, to distinguish personal census information from census information, the Census Act (1920 as amended) states:

> '"personal census information' means any census information which relates to an identifiable person or household."

The term 'identify' can be said to be reserved in legislation for the action of recognising or selecting by analysis the characteristics of a particular person or thing.

The Census Act makes it an offence to disclose any information that relates to an identifiable person or household to another person, without lawful authority. Note the term 'disclose' is reserved in legislation for the action of transferring information (identifiable or otherwise) from one party to another.

Thus the intention of the phrase in the CoP principle 5 practice 1: "… official statistics do not reveal the identity of an individual… or any private information relating to them, taking into account other relevant sources of information" is to require producers of official statistics to take account of both public and private sources of information which are deemed relevant, to consider when they are likely reasonably to be used to identify an individual.

A5. Selecting Disclosure Control Rules and Methods

The cells identified by the procedures in sections A3 and A4 as posing an unacceptable risk of disclosure are 'unsafe'. Where required, disclosure control methods can be used to reduce the risk by modifying the unsafe cells. The choice of method must balance uses to be made of the information and simplicity of approach.

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 found on the ONS Neighbourhood Statistics website (www.neighbourhood.statistics.gov.uk).

Table Design. Table redesign is recommended as a simple method that will minimise the number of unsafe cells and preserve original counts however the use of this method should be balanced against consistency in table design and publication plans.

Table 4: 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 level. City of London is combined with Hackney, Rutland UA is combined with Leicester UA and Isles of Scilly UA are combined with Cornwall UA[10]

If unsafe cells remain in the output tabulation, further protection methods should be considered in order to disguise them. If table redesign is not a feasible solution, the recommended method for post-tabular protection for most frequency tables is controlled rounding. However this method requires specialist software and therefore will not always be practical. In some cases, if the number of unsafe cells is low then cell suppression can be an alternative method. Controlled rounding and cell suppression can be implemented in the Tau-Argus software (version 3.5.0 available at http://neon.vb.cbs.nl/casc). Earlier versions are also available with 3.3.1 supported by ONS.

Cell Modification

Table 5 shows different methods which can be used to modify cell values

Table 5: 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	Original counts in the data that are not suppressed are not adjusted. Can provide protection for	Most of the information about suppressed cells will be lost. Secondary suppressions will hide	Cell suppression is used in detailed characteristics of birth tables[11]

[10] http://www.ons.gov.uk/ons/publications/re-reference-tables.html?edition=tcm%3A77-283567
[11] http://www.ons.gov.uk/ons/publications/re-reference-tables.html?edition=tcm%3A77-230095

	are called primary suppressions. To make sure that the primary suppressions cannot be derived by subtraction from totals, it may be necessary to select additional 'safe' cells for secondary suppression	zeros	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.	
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. Controlled rounding preserves the additivity of the table and can be applied to hierarchical data	Cannot be used to protect cells that are determined unsafe by a rule based on the number of statistical units contributing to a cell. 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. Random rounding requires auditing; controlled rounding requires specialist software, which is readily available.	Counts from the New Zealand Census are rounded to base 3 in their outputs[12]

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.

[12] http://www.stats.govt.nz/Census.aspx

Database Modification

Table 6: Statistical disclosure control methods - adjust the data

Method	Description	Advantages	Disadvantages	Examples
Record swapping[13]	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. Once modified all tables are produced, this can be useful when protecting online databases. Can target risky records. Gives consistent and additive tables. Counts at high geographies are unaffected.	High level of swapping may be required in order to disguise unsafe cells. Will distort distributions in the data. Method not transparent to users. It may appear as if disclosure control has not been carried out, and more metadata may be needed to support the use of this method. May be a perceived risk as cells with low counts will be published. Understanding the theory and practicality of this method may not be easy. Considerable communication and education will be required. Calculations relating to loss of data utility and doubt may need to be calculated before all output tables are produced.	Used in combination with table design to protect the 2011 Census for England, Wales, Scotland and Northern Ireland.

[13] Details of record swapping methodology can be found at http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-prospectus/new-developments-for-2011-census-results/statistical-disclosure-control/index.html

Removal of risky records	A small number of records may be unique in the data for a number of variables. Rather than protecting tables using these variables it would be simpler to remove the record	Less protection is required in the published tables without having to allow for an outlying record	A subjective decision has been made to remove information from the dataset. Users of the data may not know this has taken place or the methodology behind the removal of certain records	This is more likely to be a technique used prior to releasing microdata. There are no examples of the use of this method for ONS tabular data that have been made public.

There are many other methods of disclosure control not in the list of recommended options in the above tables. These include the method of ABS cell perturbation and over-imputation[14]. DWP have also developed a data perturbation method (Stat-Xplore) which will be used to protect many of their future releases.
An alternative option is to create a synthetic dataset which maintains all the properties of and relationships in the true dataset. From these data, non-disclosive tables can be created.

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 with limited detail in scale or providing commentaries or analytical outputs.

A6. Implementation Issues and Concerns

The proposed guidance will allow data providers to set disclosure control rules and select appropriate disclosure control methods to protect different types of published tables of statistics based on administrative sources. 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 setting disclosure control rules for tables produced from administrative data consideration ought to be given to the relationship between risk and utility. It is often impractical to aim for zero risk with the resulting outputs being of little use to prospective users of the data. Therefore any released data will have a small level of associated risk but there will also be sufficient uncertainty that any attempted identification or attribute disclosure would be correct.

There is also a relationship between disclosure risk and disclosure impact. Sufficient uncertainty will differ depending on the sensitivity of the release and consequent impact of a disclosure. This is difficult to quantify and decisions ought to be taken by those most familiar with the data. A higher level of uncertainty would be required when the table involved a

[14] Both cell perturbation and over-imputation are briefly described in this paper http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2009/wp.27.e.pdf

sensitive variable or variables, typically those of high impact. Here disclosure of any information could have a significant and longstanding effect upon both the individual and the organisation. The effect of a false claim also has to be considered. A false claim related to an individual concerning their involvement in a particular type of crime could cause great harm or distress.

When looking at published 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 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 an organisation and more widely.

Data producers should also include some (but not too much) detail on their SDC methodology within their metadata. They should also be prepared to respond to claims of disclosure, whether these claims are correct or incorrect.

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 and these should be stated in the data sharing agreements. This will ensure consistency between published statistics derived from the same source.

Any change in disclosure control rules for a published statistic raises the issue of revisions to previous releases. In general, new disclosure control rules will be implemented for future releases with the rules not being applied to past releases. An exception can be made in cases where the disclosure control rules are altered to allow more data to be released. Here it may be feasible to re-release older datasets with greater detail.

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 occurrence of an unusual infectious disease. Such decisions will only be taken at the highest level and in consultation with the Head of Profession. In many cases it will be found that these records are not statistics, but factual information and therefore subject to a different set of rules or guidance. The difference between statistics and factual information is discussed in the FOI Act and summarised in the following section.

Freedom of Information Requests
It should be noted that under the Freedom of Information Act 'statistical information' and 'factual information' are treated differently within the Section 35 exemption (the equivalent exemption in the FOISA is Section 29). Guidance on this exemption can be found on the Ministry of Justice website or, for the Scottish equivalent, on the Scottish Information Commissioner's website. 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.

Individuals have a general right of access to information held by public authorities, through the FoI or FoISA as long as this does not contravene confidentiality constraints. 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.

A7. Summary

This guidance outlines the issues concerned with protecting the confidentiality of statistics based on administrative sources 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 statistics.
If any problems arise when applying statistical disclosure control to tabular outputs from administrative data please contact
Sdc.queries@ons.gsi.gov.uk

References

Freedom of Information Act (2000)
http://www.legislation.gov.uk/ukpga/2000/36/contents

Freedom of Information (Scotland) Act 2002
http://www.legislation.gov.uk/asp/2002/13/contents

Ministry of Justice FOI website
http://www.justice.gov.uk/guidance/freedom-of-information.htm

Scottish Information Commissioner
http://www.itspublicknowledge.info

UK Statistics Authority Code of Practice for Official Statistics
http://www.statisticsauthority.gov.uk/assessment/code-of-practice/index.html

National Statistician's guidance: Confidentiality of Official Statistics
http://www.statisticsauthority.gov.uk/national-statistician/guidance/index.html

ONS (2006) Review of the Dissemination of Health Statistics – Consultation on Guidance
http://www.ons.gov.uk/ons/about-ons/consultations/closed-consultations/disclosure-review-for-health-statistics---consultation-on-guidance/disclosure-review-for-health-statistics---consultation-on-guidance.html