# Statistical Disclosure Control Protocol

## Version Control Record

<table>
<thead>
<tr>
<th>Version</th>
<th>Description of change(s)</th>
<th>Reason for change</th>
<th>Author</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>First publication</td>
<td></td>
<td>Head of Statistical Governance Team <a href="mailto:phs.statsgov@nhs.net">phs.statsgov@nhs.net</a></td>
<td>01/04/2020</td>
</tr>
</tbody>
</table>
Table of Contents

1. Introduction .................................................................................................................. 4
2. Disclosure considerations ............................................................................................. 5
   2.1 Confidentiality and Privacy ...................................................................................... 5
   2.2 Input and output SDC ............................................................................................. 6
   2.3 Attribute Disclosure ............................................................................................... 6
      2.3.1 Individual Attribute Disclosure ........................................................................ 6
      2.3.2 Group Attribute Disclosure ........................................................................... 6
   2.4 Identification and Self-Identification ..................................................................... 7
   2.5 Residual Disclosure (or “Differencing”) ............................................................... 7
      2.5.1 Differencing (To produce Small Numbers) ...................................................... 8
      2.5.2 Geographical Differencing ............................................................................ 10
   2.6 The Motivated Intruder .......................................................................................... 10
3. Disclosure Flowchart .................................................................................................... 11
   3.1 The Decision Tree ................................................................................................. 11
      3.1.1 Counts 1-9 ...................................................................................................... 11
      3.1.2 Columns dominated by zeros or 100% .......................................................... 12
      3.1.3 Population or Geography Classed as Small ............................................... 12
      3.1.4 Identifying an Individual or Gaining Additional Personal information ...... 13
   3.2 Proceeding to Disclosure Risk Assessment ........................................................... 14
4. Completing the Disclosure Risk Assessment Form ..................................................... 14
   4.1 Assessing the Likelihood of an Attempt to Disclose ............................................. 14
   4.2 Assessing the Impact of Disclosure .................................................................... 15
      4.2.1 Sensitive Outputs ......................................................................................... 16
   4.3 Calculation of the risk score ............................................................................... 17
4.4 Disclosure Control methods ..................................................................................... 17
   4.4.1 Table Redesign ............................................................................................... 18
   4.4.2 Cell Suppression .............................................................................................. 18
   4.4.3 Rounding ........................................................................................................ 19
   4.4.3 Cell (or Record) Swapping ............................................................................ 20
5. SDC for Other Statistical Outputs ............................................................................. 21
   5.1 Information Released as Restricted Management Information ............................ 21
   5.2 Information Released for Research Purposes and Tailored Analyses .................... 21
   5.3 Survey and Sampled Data .................................................................................... 22
   5.4 Microdata ............................................................................................................. 22
   5.5 Databases ............................................................................................................... 22
1. Introduction

This Statistical Disclosure Control (SDC) Protocol sets out guidance and practices for PHS staff to follow on ‘statistical disclosure control’. The protocol describes considerations of risk that should be applied when data is being released, including into the public domain e.g. publications and statistical releases, Parliamentary Questions (PQs) and all information requests including those under Freedom of Information (Scotland) Act (2002) and the Environmental information (Scotland) Regulations 2004.

This protocol is generally coherent with the guidance and accompanying best practice documents released by the Government Statistical Service which the PHS Head of Profession for Statistics has agreed that PHS follows, wherever possible.

Statistical Disclosure Control (SDC)\(^1\) is the application of methods to reduce the risk of personal identifiable information about data subjects. Methods for SDC usually restrict the amount of, or reduce, the detail of the data released. A statistical disclosure assessment, which is described in this protocol, determines whether SDC should be applied and this is a process which is necessary for documenting evidence that a standardised process of risk assessment has taken place in compliance with data protection law.

However, sometimes it is necessary to release unaltered data with conditions, such as to a NHS Board for planning and decision-making purposes, or to a safe haven for supervised access for research purposes under strict security protocols. In such cases, access is controlled by releasing them to a small number of users with conditions that they have to comply with.

SDC, isn’t always a clear and simple process to follow even when guidelines and protocols are in place.

In 2003, the UK Department of Health determined it would no longer reveal detailed information on late abortions where the number of terminations involving certain medical conditions was less than 10. This was because, prior to 2002, it released statistics which showed that one late abortion was due to cleft palate and this resulted in an outcry because the public felt that a relatively simple surgical procedure could repair cleft palates. However, in 2011, a group which opposes abortion successfully overturned, in the high court, the Department of Health’s decision not to reveal detailed information in its statistics if the numbers were less than 10.

In 2005, a member of the Scottish Parliament, through an intermediary, submitted a freedom of information request to the Common Services Agency (CSA) (i.e. NHS National Services Scotland) to supply him with the details, by census wards, of all incidents of leukaemia for both sexes, in the age range 0-14, by year, from 1990 to 2003, for all of the Dumfries and Galloway postal area. The CSA refused the request on the grounds that, for earlier years, there was a significant risk of the indirect identification of living individuals due to the low numbers resulting from the combination of the rare diagnosis, the specified age group and the small geographic area. However, upon appeal by the CSA to the Scottish

---

Information Commissioner (SIC), the SIC did agree with the CSA that privacy of children could be compromised but ordered the CSA to apply the technique of “barnadisation” which is a modification rule which adds 0, +1, or -1 to all values where the true value lies in the range from 2 to 4 and adding 0 or +1 to cells where the value is 1, and keeping 0s as 0. The CSA appealed and lost in the Scottish Court of Sessions. The CSA then appealed to the Lords of Appeal in the House of Lords who upheld the appeal and ruled that the Scottish Information Commissioner did not take due account of data protection law in respect of the personal data and sensitive personal data in concluding that the CSA should “barnadise” the data. The Lords asked both parties to work together to agree on an alternative form of disclosure control because, in “barnadisation”, the majority of cells are not modified and the risk of identification may remain unacceptably high. Since then, barnadisation has not been used as a method of statistical disclosure control in the CSA.

It should be noted that the consideration of disclosure risk may differ between publications and information requests depending on, for example, the degree of control PHS can exert on the use of the data once released. Data shared within PHS does not require SDC to be applied. However, PHS’ data protection policy must be followed in these circumstances. The person providing the data should also highlight any potentially disclosive data and, if external release of the data is planned, advise on SDC.

If confidential data are released to the public when it is not in the public interest, then this will result in loss of public trust and serious reputation damage to PHS in respect to its adherence to the “trustworthy” pillar in the Code of Practice for Statistics. There can sometimes be legal consequences for breach of confidentiality through the information commissioner’s office. Therefore, it is essential that a statistical disclosure assessment of outputs takes place before they are disseminated to a customer.

2. Disclosure considerations

2.1 Confidentiality and Privacy

In a contributing article to the National Statistician’s quality review into privacy and data confidentiality methods, it is argued that one should not assume that one has adequately protected privacy by controlling disclosure because SDC is only a mechanism for meeting confidentiality assurances. Confidentiality concerns data whereas privacy concerns individuals (or data subjects). For example, while methods of disclosure control can be applied to protect small numbers, data subjects and those who have additional information about them, may be able to recognise the data subjects within some outputs that have undergone statistical disclosure assessment and control, thus raising some individual privacy concerns. However, in many cases, confidentiality of data can indirectly and sometimes map into privacy protections.

---

2.2 Input and output SDC

Input disclosure control is control placed on the raw data prior to analysis and output SDC places controls on the analytical outputs. In PHS, due to the predominance of health and social care data collection schemes, some of which are required for analysis to produce reports on progress against national targets, performance indicators and standards, and some of which are required for high quality record linkage, input SDC is not recommended and it is necessary for the raw data to be identical to that provided by the data suppliers. Therefore, this protocol focuses on output SDC where statistical disclosure assessment is carried out on the outputs prior to release and dissemination.

2.3 Attribute Disclosure

General attribute disclosure arises when someone who has some information about a statistical unit or an individual could, with the help of data from the table, discover details that were previously unknown to them.

2.3.1 Individual Attribute Disclosure

Individual attribute disclosure arises when a data subject/individual can be identified and previously unknown information gained about them from a table.

Disclosure may arise if there is a count of 1 in a marginal row or column total of a table. For example, on examining Table 1 below, if we knew that an individual under the age of 12 in the NHS Board had received a particular treatment, we would now know that this was treatment type 1. Note that attribute disclosure can also occur from a marginal total of 2 or more, where one or more of the individuals could potentially identify information about the other and hence disclose additional information.

2.3.2 Group Attribute Disclosure

Group Attribute Disclosure arises when additional information about a certain group of people can be identified.

Disclosure can also arise from tables with larger values, where they appear in rows or columns dominated by zeros. A zero indicates that no-one in that population has that particular attribute. This can be seen in Table 1 below where all 12 to 15 year olds had treatment type 3. The risk from many zeros in a table may not be significant but in certain situations may need to be protected. Specific care should be taken if analysis shows that no one in a selected population has a particular attribute. This in itself can be disclosive about the selected population e.g. a value of zero was obtained for cancer group A in a particular NHS Board.
Table 1  Treatment Type by Age group for NHS Board X

<table>
<thead>
<tr>
<th>Treatment type</th>
<th>Age Group</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&lt;12</td>
<td>12-15</td>
<td>16-19</td>
<td>&gt;19</td>
</tr>
<tr>
<td>Type 1</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>Type 2</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>19</td>
</tr>
<tr>
<td>Type 3</td>
<td>0</td>
<td>12</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>1</td>
<td>12</td>
<td>30</td>
<td>20</td>
</tr>
</tbody>
</table>

It may also be the case whereby 100% rates are considered disclosive and is another example of group attribute disclosure. For instance, if a table were to show every girl from a particular school year in a specific board had the Human Papilloma virus (HPV) immunization, then this would provide information about each female pupil that may have previously been unknown. Although circumstances such as these may not always present information which is considered personal, PHS staff should be aware of the risks this presents and consider applying SDC.

2.4 Identification and Self-Identification

Where a table contains small cell values, particularly if there are counts of 1, more consideration is needed as identification or self-identification can lead to the discovery of rare or even unique characteristics in a population.

For certain types of information, rarity or uniqueness may encourage others to seek out the individual. The threat or reality of this could cause distress to the individual, or may lead them to claim that the statistics are inadequate to protect them, and therefore others.

For example, a table showing attendance at a drug misuse clinic by age and sex has a count of 1. 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 the result may be a lack of trust in the confidentiality of the clinic.

In order to protect against unique identification or self-identification, cells with values of 1, 2, 3 or 4 are usually considered potentially unsafe. For very sensitive outputs in small geographical areas, cell values of 5, 6, 7, 8 and 9 may additionally be considered unsafe. Although direct identification / self-identification is not necessarily a significant risk in itself, protection is often required since this could lead to attribute disclosure if other tables have been produced from the same data source and these contain additional information about an individual.

2.5 Residual Disclosure (or “Differencing”)

Residual disclosure (or differencing) occurs where outputs from the same or different sources can be combined to reveal information about an individual or a group. This can occur in a publication with many tables, for example, where the same data is cut in different ways, or from combining data from similar information requests.
For example, an enquiry from a journalist asked about the number of plastic surgery procedures carried out on teenagers under 18 years of age. A follow-up enquiry from a different journalist at the same organisation asked for information on one procedure for the age group aged under 17 years. Combining the two sets of figures provides the small number of 17 year olds who had this particular procedure.

Another example is The Drug Related Hospital Statistics (DRHS) dashboard contains data on hospitals stays, patients and new patients with a drug related diagnosis in any of the relevant diagnosis positions. The background data is available to download. There were several aggregations, e.g. NHS Board, stays per drug type, which in some cases meant that data may be disclosive (i.e. from 1 to 4 stays, patients or new patients). These were removed from the statistical programme, but in some cases there was only one instance in a table where a figure was suppressed. In these cases, the analysts chose another figure to suppress (most often the smallest remaining figure). For instance, if a figure for NHS Orkney is the only suppressed figure in a table, analysts would suppress data from another NHS Board. This would mean that the sum of NHS Boards (minus NHS Orkney) would not equal [Scotland - NHS Orkney]. In suppressing another figure, the team have made it impossible to calculate the actual number of either NHS Board.

Consideration should be given to differencing between different geographies and also changes in geographical configurations, such as in relation to the NHS Board boundary changes on 1st April 2014, and the datazone redraw on 6th November 2014 (see Annex 8).

Further problems arose with the combined dataset, in particular with stays. The sum of SMR01 and SMR04 stays should always equal the combined number of stays. The same is not true for patients and new patients, as one patient may have been admitted to an acute hospital and a psychiatric hospital in the same financial year. To preserve data confidentiality, the team cross-referenced across tables from SMR01 (inpatient), SMR04 (mental health) and combined datasets and suppressed any figures that may have been potentially disclosive. For instance, if South Lanarkshire had a suppressed figure in the mental health dataset, but not in either of the other datasets, the analysts would suppress a corresponding figure from the inpatient dataset or combined for the reasons outlined above.

2.5.1 Differencing (To produce Small Numbers)

Any sets of tables that are being released should be checked to see if they can be combined so that by inference or differencing, between rows and columns of two or more tables, disclosive cells cannot be derived. This applies to other tables produced by PHS (within the same publication or Information Request, or indeed a previous IR) or from another source (e.g. National Records of Scotland) which could be linked to the analysis that has been produced and by differencing produce numbers that may be disclosive.

When linked tables are produced from the same dataset it is not sufficient to consider the protection for each table separately. If a cell requires protection in one table, then it will require protection in all tables. Otherwise, the protection in the first table could be undone.
Tables 2A and 2B below are generated from the same dataset and provide counts for a particular characteristic by age group. The counts for 16 year olds can easily be calculated by differencing the frequencies for age bands 16-19 and 17-19. The counts for these age bands may be considered safe but the difference reveals a small and therefore potentially disclosive count for 16 years olds (one person).

Table 2A

<table>
<thead>
<tr>
<th>Age</th>
<th>&lt; 16</th>
<th>16-19</th>
<th>20-24</th>
<th>≥ 25</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>5</td>
<td>26</td>
<td>13</td>
<td>16</td>
<td>60</td>
</tr>
</tbody>
</table>

Table 2B

<table>
<thead>
<tr>
<th>Age</th>
<th>&lt; 17</th>
<th>17-19</th>
<th>20-24</th>
<th>≥ 25</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>6</td>
<td>25</td>
<td>13</td>
<td>16</td>
<td>60</td>
</tr>
</tbody>
</table>

It may also be possible to derive disclosive cells from information in one table. For example, in a recent table on abortions it was possible to derive that a group of eight women had had a surgical termination at less than ten weeks in a particular NHS board, from the overall total and the percentage breakdowns of the estimated gestation and method of termination.

Where tables provide data in terms of rates or percentages, the figures themselves may not be disclosive. However, if the rate or percentage is based on an unsafe cell and it is possible by linking with other tables to recover the original count, then the cell with the rate or percentage is itself unsafe.

Some protection can be provided by the use of rounding rates, or percentages but care still needs to be taken to avoid disclosure. Protection will be provided if the base population from which the rate or percentage is calculated is sufficiently large, since the implied count could be a range of values (however this range must be large enough to satisfy disclosure rules and thresholds).

It is also important to consider if denominators are easily or publicly available (from another table in the publication or perhaps populations which can be obtained from the NRS website). If they are, then cells must be considered unsafe and SDC should be applied. If denominators are not known to be available from another source, then figures can be considered safe. Note that crude rates may need to be handled differently, considering the ease of recalculating original counts.

It may, however, be safe to publish directly and indirectly standardised rates. Calculating original counts from such rates would generally require a degree of specialised technical expertise in the associated field which could lower the likelihood of anyone attempting or successfully attempting to calculate the exact numbers. In such cases other factors that should be considered are the availability of population counts, the time period involved (are data aggregated over a number of years) and any other known/unknown variables that may
be required to calculate original figures. Each case should be examined individually before deciding if some form of SDC is necessary.

2.5.2 Geographical Differencing

Suppose that figures are produced for the two geographical areas A and B as shown in Figure 1, where A is a subset of B. Data for geographical area C could easily be produced by subtraction (differencing). If two tables are produced for different geographies from the same dataset, then disclosure by differencing can occur even if the two tables have been protected independently.

Figure 1  Geographic Differencing Illustration

Consideration should be given when releasing data at different geographical levels in PHS, as many of the geographies reported on are not coterminous (for example West Lothian council area had an overlap with Lanarkshire NHS Board using the 2006 NHS Board boundaries).

Data can be provided at a geographical organisational level (i.e. NHS Board, Local Authority) and also at a physical geography level (i.e. postcode sector, datazone or intermediate zone levels). Very few of these areas have a one-to-one mapping and often the use of “best fit” geography is required when ‘building’ a higher geographical area from small areas. This does not always provide exactly the same physical geographical region as the true region and therefore differences will occur in the data reported.

Annex B contains further information on geographies and populations, including details of the most recent geographical changes and the impact this has regarding disclosure.

2.6 The Motivated Intruder

When releasing data, it should be borne in mind that our data could be combined with that from other local sources to identify individual(s) and disclose further personal details about them.

This situation may arise when small cell values are presented for small geographies. In larger populations, the effort and expertise required to discover more details about an individual may be considered disproportionate, but when the base population is decreased (for example consider data for small geographies such as council area or an Island Board), it will, in many cases, become easier to find additional information about individuals.
Although locally sourced data may reveal the identity of an individual, it may be NSS’s publication that prompts the motivated intruder to start an investigation. A motivated intruder is someone who deliberately tries to gain information about some person or business e.g. potentially the media or ‘nosy neighbour’. It may not always be necessary, or feasible for PHS to consider all local sources of data, but it is necessary to consider the information likely to be available to third parties, and assess the likelihood and risk of an intruder being motivated enough to track down individual(s). It must be noted that under section 171 of the Data Protection Act 2018, it is a criminal offence for a person knowingly or recklessly to re-identify information that is de-identified personal data without the consent of the controller responsible for de-identifying the personal data.

3. Disclosure Flowchart

3.1 The Decision Tree

This protocol sets out ‘guidelines’ for risk assessment of disclosure arising from a statistical release. It is important to note that this protocol does not set out a particular formula that provides a measure of risk for every scenario. Rather, the emphasis is on the need for judgement to be made, on a case-by-case basis, of the risk and provides guidance on how best to assess the risk. The decision to proceed to completing a disclosure risk assessment form is dependent on whether any of the following is true of the output:

- Counts 1-9
- Columns and rows dominated by zeros (or 100% rates)
- Population or geography classed as small
- An individual can be identified or additional personal or sensitive information about an individual can be gained from the data

The decision tree is shown in the statistical disclosure control flowchart shown in Annex D. The flowchart has been designed to help PHS staff assess the risk of disclosure and decide on whether disclosure control is necessary. When assessing the risks of disclosure in data for management information purposes the same considerations will apply as for published data or data released to a customer.

3.1.1 Counts 1-9

Where the outputs include counts of 1 to 9, the risk assessment form shown in Annex E should be completed. This is because it has been determined in PHS that there are two levels of consideration of small numbers: 1-4 across for all outputs and 5-9 where the population is small and/or the outputs are sensitive and/or the columns are dominated by zeros. In addition, where the count is associated with 1 or 2 practitioners, or 1 or 2 named locations, the disclosure assessment form must be completed. For example, the data shown

---

might not identify individual(s) receiving a certain type of care for a certain condition, but
might help identify individual(s) providing certain types of care. A scenario might be where a
particular surgical operation is carried out by only one clinician (or in only one hospital) in a
certain NHS Board – release of such data might disclose information particular to the
individual clinician (or institution), particularly if it is of a sensitive nature.

3.1.2 Columns dominated by zeros or 100%

Where this is the case in the outputs, the risk assessment form should be completed. This is
due to the risk of group attribute disclosure as described in section 2.3.2.

3.1.3 Population or Geography Classed as Small

Small populations and geographies increase the likelihood of disclosure. For purposes of
SDC within PHS, the following guidance is given for ‘small’ population and geography:

- **‘small population’** – it is emphasised that there is no definitive threshold below which a
  population can be considered to be small. The population threshold should be dependent on
  the situation and should be based on the identifiable population at risk. In general, if the
  population at risk is more than 5,000 then the likelihood of disclosure is considered to be
  low. However, you should still assess the population at risk for each output, taking into
  account factors that affect the likelihood of disclosure such as sensitivity, geography (rural
  versus urban), etc.

  Where the population at risk is smaller than 5,000 a more detailed assessment of the
  likelihood of disclosure should be undertaken to determine a minimum population threshold
  adequate to provide protection. This should take account of the scenarios described in
  Sections 2.3 to 2.5 by which individuals could be identified. For example, this might involve
  considering the minimum number of households or schools to deter an intruder from trying
  to identify an individual household or school or the effect on individuals of statistics being
  released in local media.

  The ‘population at risk’ is the denominator for a cell in question; this may not be the entire
  population for a particular table or a particular cell – an example would be where numbers
  of births are shown for ‘<20’, the population at risk would be ‘females aged 15-19’ not
  ‘females aged <20’). In some cases, the base population itself may require discussion and
  judgement but commonly will be based on age and sex.

- **‘small geography’** – the considerations for geography are similar to those for population and
  very often the two will be inter-linked. However, the sparsity of the population within an
  area could be a factor affecting the risk of disclosure in some cases. In general, a small
  geography would mean data presented for individual island NHS Boards, Local Authorities or
  data presented below NHS Board level.

- **Institution level** - data presented at hospital level or below (e.g. clinics, GP practices) can
  increase the risk of disclosure of patients.
3.1.4 Identifying an Individual or Gaining Additional Personal information

Where it is obvious from the output that personal data about a specific individual or a group of individuals is revealed or, in some cases, where it is already known who the individual is, but additional personal or sensitive information is revealed in the output data, then the disclosure risk assessment form should be completed which may direct the staff to applying disclosure control. One example is where an individual has already identified themselves in the public domain as the only person with a rare condition in Scotland, it is essential to ensure that no additional personal or sensitive information about that individual is released by PHS which could result in the public gaining more knowledge about the individual as a result of data released by PHS.

Identifying individuals and gaining additional personal or sensitive information: Table 3 shows that there are 625 women from NHS Board X that had an abortion, and from this table we learn that all of these were on ‘Ground C’ (risks of injury to the physical and mental health of the <24 week pregnant woman). So therefore we are gaining additional sensitive or personal information about these individuals.

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Abortions by NHS Board, age and grounds</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NHS Board</td>
</tr>
<tr>
<td></td>
<td>X</td>
</tr>
<tr>
<td>All Abortions</td>
<td>625</td>
</tr>
<tr>
<td>Rate per 1000 live births</td>
<td>183.6</td>
</tr>
<tr>
<td>Rate per 1000 women aged 15-44</td>
<td>10.6</td>
</tr>
<tr>
<td>Age of Woman</td>
<td></td>
</tr>
<tr>
<td>Under 20</td>
<td>174</td>
</tr>
<tr>
<td>20 - 24</td>
<td>174</td>
</tr>
<tr>
<td>25 - 29</td>
<td>121</td>
</tr>
<tr>
<td>30 - 34</td>
<td>73</td>
</tr>
<tr>
<td>35 - 39</td>
<td>50</td>
</tr>
<tr>
<td>40+</td>
<td>33</td>
</tr>
<tr>
<td>Grounds for abortion</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>625</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
</tr>
<tr>
<td>Emergency</td>
<td>0</td>
</tr>
</tbody>
</table>

Identifying individuals and not gaining any additional personal or sensitive information: Table 4 shows the type of vouchers of General Ophthalmic Services we can see that there was one bifocal complex payment on the Western Isles but we do not learn anything additional about this person.
Table 4  NHS vouchers GOS(S)3 by type and NHS board
Year ending 31 March 2008

<table>
<thead>
<tr>
<th>Number</th>
<th>Scotland</th>
<th>Orkney</th>
<th>Shetland</th>
<th>Tayside</th>
<th>Western Isles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single vision</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>305,366</td>
<td>602</td>
<td>1,088</td>
<td>19,855</td>
<td>1,521</td>
</tr>
<tr>
<td>B</td>
<td>56,444</td>
<td>70</td>
<td>170</td>
<td>3,273</td>
<td>204</td>
</tr>
<tr>
<td>C</td>
<td>3,410</td>
<td>3</td>
<td>5</td>
<td>193</td>
<td>9</td>
</tr>
<tr>
<td>D</td>
<td>1,706</td>
<td>2</td>
<td>5</td>
<td>104</td>
<td>5</td>
</tr>
<tr>
<td>Bifocal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>67,439</td>
<td>71</td>
<td>140</td>
<td>5,160</td>
<td>270</td>
</tr>
<tr>
<td>F</td>
<td>13,719</td>
<td>4</td>
<td>24</td>
<td>1,031</td>
<td>36</td>
</tr>
<tr>
<td>G</td>
<td>373</td>
<td>0</td>
<td>1</td>
<td>30</td>
<td>2</td>
</tr>
<tr>
<td>H</td>
<td>279</td>
<td>0</td>
<td>0</td>
<td>24</td>
<td>0</td>
</tr>
<tr>
<td>Complex payment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single</td>
<td>1,403</td>
<td>2</td>
<td>0</td>
<td>142</td>
<td>5</td>
</tr>
<tr>
<td>Bifocal</td>
<td>577</td>
<td>1</td>
<td>0</td>
<td>65</td>
<td>1</td>
</tr>
<tr>
<td>All voucher types</td>
<td>450,716</td>
<td>755</td>
<td>1,433</td>
<td>29,877</td>
<td>2,053</td>
</tr>
</tbody>
</table>

3.2 Proceeding to Disclosure Risk Assessment

As illustrated in Annex D, having determined whether any of the 4 sets of conditions apply to the output data, if the response to everyone of them is ‘no’, then the output can be released without completing a risk assessment form. However, if the response is ‘yes’ to any of the 4 conditions, then a disclosure risk assessment form must be completed. Both parts one and two of the Disclosure Risk Assessment Form must be completed.

4. Completing the Disclosure Risk Assessment Form

The first part of the risk assessment form in Annex E requires the staff to calculate the risk assessment score which is the product of a score denoting the likelihood of an attempt to disclose and a score denoting the impact of disclosure.

The risk score is based on the following definitions of low (1), medium (2) and high (3) risk. It is not always a cut-and-dry process and judgement would sometimes need to be made. It is advisable for the assessor to discuss their assessment with their team and if consensus is not gained, the Statistical Governance Team must be consulted at phs.statsgov@nhs.net for advice. Wherever possible, discussion should take place with the customer, without revealing the detailed methodology, to decide on the most appropriate SDC method to protect confidentiality.

4.1 Assessing the Likelihood of an Attempt to Disclose

Assessing the likelihood of an attempt to disclose requires the assessor to take the following into consideration:

- Who wants the data? – e.g. Named contact, position/role, organization, policy makers, non-NHS partner organizations, NHS Steering Group
  - How will the data be accessed? – e.g. controlled, public domain, media
- What data do they want? – e.g. data sources, variables, time periods, etc.
- Why do they want the data? – e.g. inform committee meeting, FOI, PQ, research paper, publication, SG policy, inform healthcare decisions
- Is there similar known data out there which could be combined with what you are about to release? – e.g. newspaper articles, internet, previously published information

For example, the Head of Information Services of a local NHS Board requiring data about patients in their NHS Board area for operational purposes within their Board in relation to numbers of patients waiting over 9 months for their surgical procedure will imply that there is a low likelihood of an attempt to disclose. Therefore, the score will be 1 (low). However, a non-core customer with a personalised non-organisation email address requiring the same information with no plausible explanation as to why data at that disaggregated level are required would imply that there is a high risk of a likelihood of an attempt to disclose and the score will be 3 (high).

Another consideration is the data that the customer wants. Low risk of a likelihood of an attempt to disclose for any customer would mean that it would be difficult for someone to identify disclosive information from the release. This would be the case where there is little chance of differencing between tables as there are not many tables released from the dataset or if the topic was not sensitive, the cell value was high and presented at Scotland level. Releases could also be considered low risk if the data is from a large population. However, this would depend on the variables used in the released tables.

Yet another consideration is whether it is known that there is data out there in the public domain which could potentially be searched for and used to combine with the relatively safe data that the assessor is planning to release. In such cases, for those customers who may not have an operational care-related reason for requesting the data, the risk may be high (3) or medium (2) depending on the sensitivity and age of the data.

Another consideration is the type of data being requested. The likelihood of an attempt to disclose can be affected by the timeline associated with the data. For instance, if the information is aggregated over a number of years rather than single years then this could lower the risk involved as identifying individuals may be more difficult. This may also be the case for information which is not particularly recent and so the lapse in time from the period in which the figures are based to when the data is released should also be a consideration when determining the risk.

4.2 Assessing the Impact of Disclosure

The assessment of the impact of disclosure can be considered to be somewhat subjective. The impact of disclosure is the consideration of what would happen if the data being released to a customer, without disclosure controls, were released to persons not authorised to receive it, e.g. the public. The assessment should be undertaken by a member of staff with experience of the information being released. The assessment of impact should take into account those who may have an interest in the data being released, the views of patients and carers and potentially disclosive situations which could occur through disclosure.

When assessing the impact of disclosure, the following should be taken into consideration:
• Is the output sensitive? – see section 4.2.1
• What data do they want? - Data sources, variables, time periods, etc.
• Will additional personal or sensitive information be gained about an individual? e.g. geographical area of residence, age, previous diagnosis

Sensitive outputs are discussed in the section following. However, it should be noted that while a topic may be sensitive, e.g. drug and alcohol treatment, the output may not be sensitive because it is about waiting times of drug and alcohol treatment centres presented as numbers of weeks broken down by NHS Board or treatment centre. Therefore, for this example, the impact of disclosure would be low (1).

The type of data requested is another consideration for the impact of disclosure. Releasing numbers of patients having dermatological outpatient procedures broken down by location and consultant has a low (1) impact of disclosure; whereas, small numbers having treatment for a specific psychiatric disorder broken down by location (including within Island Boards), age group, sex and consultant will have a high (3) impact of disclosure.

Sometimes it is widely known that only one (unknown) individual has a very rare condition. Releasing data in response to this type of condition broken down by numbers of patients, Board of treatment, Board of residence, age group, and sex would result in additional personal information being released about an individual which will result in a high impact of disclosure.

Not all scenarios classed as sensitive will necessarily be scored as high risk. Examples may be where there are specific local issues or where the topic concerned is currently the subject of particular media attention. However, the assessor must discuss any unusual scenarios with their team.

4.2.1 Sensitive Outputs

Analysis involving some sensitive topics may result in outputs which are sensitive or outputs which are not sensitive. For example, when sensitive topics are presented as waiting times or average length of stay, the outputs may not necessarily be sensitive.

Although no definitive list exists, as a guide, the following are usually classed as sensitive topics or vulnerable populations where care needs to be taken when calculating the risk assessment score:
• Sexually transmitted disease
• Abortion
• Suicide, self-harm
• Drug and alcohol misuse
• Mental health
• Contraceptive prescriptions
• Vulnerable populations e.g. people subject to the Looked After Children (Scotland) Regulations 2009, people subject to Adult Support & Protection (Scotland) Act 2007

Increasingly, with personalised medicine advancements increasing at pace, it is known that genetics remains just one of several factors that contribute to people's risk of developing most common diseases, including diet, lifestyle, and environmental exposures. Therefore,
this emerging and growing area of human genomics and genome sequencing is also classed as sensitive.

4.3 Calculation of the risk score

Having obtained risk scores for the ‘likelihood’ and ‘impact’ in terms of ‘high’ (risk score = 3), ‘medium’ (score = 2) or ‘low’ (score = 1), the scores for ‘likelihood’ and ‘impact’ are then multiplied.

Where the resulting score is 4 or greater then disclosure control methods should be applied. In addition, if it is considered that sensitive information about an individual exists within the output, then SDC may be considered appropriate where the risk assessment score is 3 or less.

The risk score will be particularly important for some scenarios where the score will impact on the decision on whether to release or to protect data.

4.4 Disclosure Control methods

Where the need for SDC is identified then there are a range of methods that can be considered for use. The choice of method should balance uses to be made of the information and simplicity of approach. These methods can be divided into three categories:

- those that determine the design of the table – table redesign
- those that modify the values in the table – cell suppression, cell swapping, controlled rounding
- those that adjust the data before tables are designed – used in open data releases and they include cell suppression, cell swapping, controlled rounding

Alternative methods for presenting data can be considered as an approach for providing users access to information without disclosing the underlying data. In some cases, this will provide a more robust analysis than reliance on the accuracy of small cell values – for example, these could include presenting data graphically with limited detail in scale.

In applying certain types of SDC, programmes may have been written in R or Python by the user communities in PHS. Please contact the Statistical Governance Team for more information.

Where SDC is required, firstly consider the following in order of preference:

- Table redesign – e.g. grouping or aggregating cells
- Cell Suppression – applying primary and secondary suppression to unsafe cells and replacing their values with ‘*’ in tables or using the qualifier ‘c’ in open data
- Rounding – applying controlled rounding to a multiple of a set base such as 3 or 5 up to a maximum value – must have prior discussion with the Statistical Governance Team
- Cell swapping – targeted or random cell swapping methods – must have prior discussion with the Statistical Governance Team

Where high risk outputs need to be released to a specified table design, the Statistical Governance Team may advice that a combination of the above methods be used as SDC.
The SDC methods presented here are consistent with methods proposed by the Government Statistical Service\textsuperscript{7}.

4.4.1 Table Redesign

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

Description: Remove unsafe cells by, for example:
- 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
- clustering values for the Island Boards into a single ‘Island Board’ descriptor

Advantages:
- original counts in the data are not damaged
- easy to implement
- easily understood by user

Disadvantages:
- detail in the table will be reduced
- may be policy or practical reasons for requiring a particular table design

Annex C provides further information and example.

Depending on the nature of the request and/or certain circumstances, PHS staff may wish to explain that the design of the table has been influenced by disclosure. For example, table presentation may differ from that contained in previous publications. In this instance the following could be used, amending as required.

“The design of a number of the tables presented in this publication has been revised from previous editions. These changes attempt to minimise the risk of disclosure and to help maintain confidentiality.”

4.4.2 Cell Suppression

Description: Unsafe cells are not released. They are suppressed and replaced by a ’*’ (an asterisk) 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.

Advantages:
- original counts in the data that are not suppressed are not adjusted
- can provide protection for zeros

allows original/requested structure to be maintained
depending on number of cells ‘at risk’, can be preferable to table re-design

Disadvantages:
- most of the information about suppressed cells will be lost
- secondary suppressions will hide information in safe cells (this could include totals)
- information loss may be high if more than a few suppressions are required
- any potentially disclosive zeros would need to be suppressed
- does not always protect against disclosure by differencing

Past experience has shown that it is good practice to present tables with totals. If totals are not included, then a customer could return to ask for totals. This must then be considered in conjunction with any cell suppression applied to the original table and may result in some totals being suppressed to ensure previously suppressed figures cannot be calculated through differencing.

The comparison of data from numerous tables must also be considered (including previously released data) to ensure protection against differencing and so suppressing data can be time consuming and complicated.

The following rules should be applied for suppression (primary and secondary):
- replace both primary and secondary suppressed cells with ‘*’ (an asterisk). This symbol should not be used for any other value. Do not use different symbols for primary and secondary suppressions and do not indicate which cell are secondary suppressed.
- Values of zero should not automatically be selected for primary suppression. On some occasions suppression of zeros may be required for secondary suppression or where rows and/or columns are dominated by zeros.
- Care must be taken with any secondary suppression of data. Normally the next smallest number would be selected for secondary suppression. However, this is not always the best option. Selecting another larger number may lead to less cell suppression within the table, thereby maximising utility.
- If only 1s, 2s, 3s and 4s have been suppressed and there are marginal totals then secondary suppression may be necessary to shield their value.
- A footnote advising that cells have been suppressed should be added to all relevant tables (and not only specified in an attached email, for example) and be consistent for all tables within a publication. The footnote should not detail the values suppressed e.g. <5 or <10. The example below can be used.

* Indicates values that have been suppressed due to the potential risk of disclosure and to help maintain confidentiality.

See Annex C for further information and example.

4.4.3 Rounding

There are a range of methods of applying rounding as a method of SDC, including: controlled rounding, deterministic rounding and random rounding. If rounding is to be used the method currently recommended for PHS is controlled rounding where the external cell totals equal the sum of their internally rounded cells (i.e. no loss of additivity).
Description: Involves adjusting the values in all cells in a table to a specified base, so as to create uncertainty about the real value for any cell, while adding a small but acceptable amount of distortion to the data. The base for rounding can be chosen with common choices being 3 or 5. All rounded values (other than zeros) are then integer multiples of 3 or 5 respectively.

Advantages:
- if the number of unsafe cells is large then the table can be protected while still providing counts for all cells and without altering the design of the table
- will protect zeroes without removing them since, within a table rounded to base 5 for example, a zero could represent any count between 0 and 4
- cells rounded to a common base in such a way as to preserve additivity to totals within table (unlike random rounding where all figures including totals are rounded randomly and so may not be additive)
- fully protects against disclosure by differencing

Disadvantages:
- difficulties in disguising cells in which the count can be associated with either 1 or 2 practitioners/hospitals whom it may be necessary to protect
- if user requires exact counts rounded values would not be appropriate
- if population size is small then rounding may not offer enough protection against identification
- can at times distort data to such a degree that original trends cannot be identified. Care must therefore be taken to avoid this whilst ensuring trends that do not actually exist cannot be wrongly interpreted
- may be prone to effects of data revisions, for example updates to figures that are contained in future editions of a publication series may require a different pattern of rounding than that used in previous presentation of the figures
- may not be helpful to those users of PHS’s statistics who are familiar with historical pre-PHS methods of presentation (including SDC)
- totals may be adjusted, thereby altering ‘headline’ figures

Due to these various issues, rounding is currently not a first preference method of SDC within PHS and should not be applied without prior discussion with the PHS Statistical Governance Team. Any use of rounding should be carefully considered. For example, there may be an impact on the use and interpretation of any rounded figures. It is therefore essential to ensure that any information provided is not unnecessarily misleading to the user.

4.4.3 Cell (or Record) Swapping

Cell swapping is used extensively in some other organisations for modifying input data relating to non-health datasets. However, in PHS, this is normally a method of last resort for output data which must be discussed with the Statistical Governance Team as it is not applicable to many scenarios.

Record swapping introduces an element of doubt into the dataset by swapping a small number of records that contain similar characteristics. While the number of records being
swapped is small, the user is now unable to identify whether a record has been swapped or not.

Random cell swapping maintains a higher data utility compared with targeted cell swapping at the same swap rate. However, targeted cell swapping provides a greater level of protection against disclosure since it targets the risky cells.

Please discuss with the Statistical Governance Team before using this method.

**Advantages:**
- Protects against disclosure by differencing
- Swapping rates are flexible
- Can target risky records
- Can be applied to the base data (microdata)
- Tables can be made to be additive
- Consistent totals between tables
- Counts at high geographies are unaffected

**Disadvantages:**
- Table not visibly perturbed – so users need clear explanations to ensure transparency
- Public may perceive that no SDC has been applied
- May introduce some bias by distorting distributions in the data
- High level of swapping may be required in order to disguise unsafe cells

# 5. SDC for Other Statistical Outputs

## 5.1 Information Released as Restricted Management Information

PHS has a duty, set out in legislation to provide Scottish Ministers with services that support the prevention of illness, the care of persons suffering from illness, and the after-care of such persons. In addition, PHS has a duty to cooperate with other special health boards, local health boards, NHS National Services Scotland and Healthcare Improvement Scotland in order to promote the improvement of the physical and mental health of the people of Scotland.

In discharging these duties, PHS provides restricted management information to the Scottish Government and health Boards in support of the proper management of the health service. PHS facilitates this in a way that does not compromise its duties under data protection rules. Annex A contains guidance on handling requests for information for management purposes and also information for data quality assurance purposes.

## 5.2 Information Released for Research Purposes and Tailored Analyses

Current requests for the release of personal information in support of research may require an application to be made to the Public Benefit and Privacy Panel for Health and Social Care where the application will undergo proportionate governance and independent scrutiny by
patient representatives, Caldicott Guardian representatives, and information governance specialists. Further details can be found here: https://www.informationgovernance.scot.nhs.uk/pbpphsc/

For other requests, contact the PHS data protection team at phs.dataprotection@nhs.net.

5.3 Survey and Sampled Data

When considering disclosure control for reporting results produced from survey data, there is no set protocol that covers all outputs. It should be acknowledged that disclosure risk must be reviewed on a case by case basis. Examples of points to consider include:

- The topic of the survey and any sensitivities associated with the responses
- Any commitments / assurances that were made to survey respondents in respect of how the results would be reported
- The sample size of the overall survey and any results published at sub sample level (e.g. gender, age group, designations, geography)
- When reporting free text responses, is it possible to identify individuals?
- Survey results may require restricted outputs for quality purposes e.g. avoid reporting estimates with potentially high sampling errors

Where surveys are run by PHS on behalf of an external organisation, consultation in relation to disclosure protocol should be discussed with the customer.

5.4 Microdata

Micro-data is data collected at an individual level, such as patient level data or data collected for a survey or for the Census.

The disclosure protocol does not apply for micro-data as different processes apply and an agreement such as a data processing agreement or a completed data release and linkage form is required.

For more information on the release of micro-data, please contact the Statistical Governance team at phs.statsgov@nhs.net.

5.5. Databases

These concern access to databases with personal information such as PRISMS, ACaDMe, and Outpatients. This guidance refers to ‘databases’ (or ‘warehouses’, or ‘datamarts’, etc.) that are made accessible to PHS and non-PHS staff for analysis e.g. the extraction of reports. There will be a range of types of databases made available to non-NSS staff and for some (e.g. ACADME, PRISMS) there may be wider confidentiality issues. It is important that confidentiality arrangements that are specific to individual databases are followed.

A formal procedure for authorising access to each database exists in PHS. The decision on whether an individual requires access is the responsibility of the organisation (e.g. NHS Board) accessing the database. These organisations – in conjunction with PHS - are responsible for ensuring that only appropriate staff have access to the database. They must
also make employees fully aware of their responsibilities in relation to disclosure control and ensure they adhere to confidentiality principles and comply with data protection law obligations.

The PHS information asset owner is ultimately responsible for authorising access to the database they control under established authentication and access protocols in PHS. Information made available in these types of formats might commonly be considered management information for restricted access.

5.6 Open Data

Data ready for open data release must go through the disclosure assessment process described in this protocol. For open data, depending on the dataset, when sensitive outputs at relatively low granular level is being prepared for release, it may sometimes be necessary to consider cells with values less than 20 as being unsafe and apply rounding as an SDC method. Staff should consult the special instructions in the PHS guidance for creating and releasing open data which should be available in the intranet or by emailing and seeking advice from the Statistical Governance Team at phs.statsgov@nhs.net.

5.7 Synthetic Data

Synthetic data is sometimes viewed as an alternative to SDC but with a similar goal to release a useful dataset whilst maintaining data subject confidentiality. Synthetic data is also sometimes referred to as ‘fake data’ when it is developed from scratch and determined to have no disclosure issues due to the method of creation of a new dataset which has similar structures to a real dataset. Developments in PHS have focused on using non-parametric methods to develop synthetic data such as Machine Learning techniques which are better at capturing non-linear relationships for generating synthetic data. The techniques include artificial neural network approximate functions, variational auto encoder and generative adversarial networks. It is expected that rigorously tested synthetic data will become available in PHS in due course for users who would like to test what a dataset can do before applying for access to specific variables in the real data.

6. Documenting Disclosure Decisions

The issue of disclosure concerns all of PHS’s outputs (including publications and Information Requests). To document decisions taken on disclosure issues a Disclosure Risk Assessment Form should be completed.

A Disclosure Risk Assessment Form should be completed for:

- all PHS publications (including contributions to ‘non PHS’ publications)
- certain Information Requests, including FOIs and PQs (see Section 6.1 below)

---

Note that information released by PHS in response to a PQ should be treated as being publicly released and should always be subject to our own guidance and policies. SDC should therefore always be considered before releasing information in response to a PQ.

A Disclosure Risk Assessment Form should be completed when the disclosure flowchart has instructed to undertake a risk assessment (with the resultant outcome being to either protect or release).

(Section 6.1 below describes particular situations where this may not be the case.)

6.1 When to Complete a Disclosure Form for Information Requests (IRs)

A Disclosure Risk Assessment Form should be completed for all IRs where the end point by using the Disclosure Flowchart is “Protection Required”. You should also complete a Disclosure Risk Assessment Form if the Disclosure Flowchart instructs you to undertake a “Risk Assessment”. Note that where data relates solely to the customer’s own organisation or limited other situations e.g. groups with national responsibilities, whether the risk score directs you to protect or not, you should document your decisions in the Disclosure Risk Assessment Form and ensure the standard text in Annex A is attached to the data.

You should however not complete a Disclosure Risk Assessment Form where you are instructed using the Disclosure Flowchart to “release” data.

Prior to the release of information, where possible, the appropriateness of the customer requesting and receiving the information should always be checked to ensure that it is appropriate for them to receive the data, including without any SDC applied. It should also be noted that data contained in Information Requests should be released in compliance with data protection law.

6.2 Part One: Risk Assessment

A Risk Assessment should only be undertaken when instructed on the Disclosure Flowchart.

The risk assessment involves assigning a score of Low (1), Medium (2) or High (3) to the ‘likelihood of an attempt of disclosure’ and the ‘impact of any disclosure’. The risk assessment may involve discussion with colleagues and managers within your team, and the Statistical Governance team.

If by multiplying the ‘likelihood’ by ‘impact’ the score is 4 or above, then some form of disclosure control should be applied to the data prior to release. If the risk score is less than 4, then the data can be released however you should complete the relevant questions in Part 2 of this form. In some circumstances, it may be judged that disclosure control should be applied when the risk score is 3 (for example if sensitive information about an individual or group of individuals is in the output table).

If a risk assessment has been carried out, then Part Two of the disclosure form should also be completed.
6.3 Part Two: Details of Disclosure Control

Part Two of the disclosure form should be completed when the disclosure flowchart has instructed to protect or undertake a risk assessment (following completion of Part One).

**Question 1:** Did you apply any disclosure control techniques prior to releasing the data?

**Question 2:** Indicate all of the potential disclosure risks with the data e.g. sensitive output, counts of 1-4, etc.

**Question 3:** Indicate which disclosure control techniques you applied?

**Question 4:** Describe the effect on the output of applying the disclosure control techniques. An example might be:

‘All cells values of 1-4 in Tables 1 to 10 of this publication were suppressed because these values were based on a ‘small’ population. Secondary suppression was also required within these tables to ensure primary suppressed values could not be calculated. Within Tables 11 to 15, age groups were combined to aggregate small numbers.’

Provide as much detail as possible here. This information will be used as a reference for future publications or similar IRs and also for reviewing past SDC decisions.

**Question 5:** Additional comments. This space should be used to document any other information you feel appropriate and also to document why you did not decide to apply any disclosure control prior to release. You may also include outcome of discussions if it was a complex one which was discussed with the Statistical Governance Team.

The completed disclosure form should be passed to a responsible person who is familiar with the subject matter for sign-off. This person could be a service manager, information consultant, team manager, or equivalent, or their authorised deputy.

Sign-off should be done electronically and the form then stored by the PHS staff in a shared directory for future reference. **Only once the Disclosure form is signed off can the data be released.**

Completed Disclosure Risk Assessment Form provide evidence for future assessment of PHS’s practice and evolution of guidance. It is also a good reference for future releases and evidence for any investigation if a customer challenges the level of disclosure control applied or complains to the Information Commissioner’s Office (ICO).
7. Annex A - Information released for Management or Data Quality Assurance Purposes (including Peer Review)

If handling a request for management information or information for data quality assurance purposes users must also use the Disclosure Flowchart from where they may or may not be directed to proceed to complete a risk assessment form, depending on the risk score. Following this, the decision may in some cases result in a different outcome to that of the flowchart. This is because of the type and role of the recipients of the data.

7.1 Information released for Management Purposes

Management information is information provided for management (or ‘operational’ or ‘clinical management’) purposes. One of PHS’s key functions is to assist partner organisations (mainly NHS Boards, but also others) plan, monitor and evaluate health and care services; this often requires supporting data. Sometimes this data would not be deemed suitable for general publication so PHS needs to minimise the risks associated with this data while still providing partner organisations with the data they need to operate, plan and monitor services.

As for published data, the risks of statistical disclosure still need to be considered, and the principles involved in risk assessment are the same. However, in general, given the recipients and their intended use of the data, the risks of disclosure will be much lower than for general publications, in particular the risk of an attempt to disclose. Risks can often be reduced by taking certain steps such as marking the information released as management information only and discussing the risks of disclosure with the recipient. This is what allows data, that would be deemed potentially disclosive in the public domain, to be released to the health service and key partners.

Often the greatest risk that needs to be considered is not the risks of disclosure by the recipients themselves but the risks posed by onward distribution or inadvertent release into the public domain. Focus therefore will often be on the actions necessary to reduce this risk.

While an exact definition of restricted ‘management information’ is not possible, the following list should be taken as a guide:

- data to monitor and improve a service;
- reports and analysis provided to programme boards or steering groups;
- data to monitor targets;
- service planning
- quality control/improvement
- performance management/targets
- routine reports to NHS Boards etc.;

Often these purposes will require data that may contain small numbers or be disclosive in other ways, even within just a small proportion of the outputs. Sometimes the user requires the full unadjusted output, due, for example, to the inclusion of data for small NHS Boards or for tracking financial flows or target performance. Often management information will be in the form of routine standard reports which could, on occasion, contain some cells with a higher risk of disclosure due to small numbers.
A key principle however, as for published data, is to release data only at the level of detail needed to fit the purposes of the user and avoid releasing unnecessary disaggregations that may pose a higher risk of disclosure. It is important to discuss needs with the user when the request involves potentially disclosive data.

Regardless of who is requesting the information or the intended audience of the data, if the decision has been made to release potentially disclosive data for management information purposes it is PHS’s responsibility to highlight this to the recipient and the following rules should be applied:

- Each table within the release should be labelled as ‘Management Information only, not for onward distribution’. This statement should appear prominently at the top of each table.
- The **standard text** below should be attached as a footnote to each table:
  
  *This information has been released for management information purposes only. The data have not been adjusted to protect against potential disclosure risks and may contain information which enables (perhaps with the aid of further knowledge of the topic) an individual patient or member of staff to be identified. Please ensure circulation is restricted and that patient confidentiality is not compromised. Please contact phs.statsgov@nhs.net if you have any queries around this or if there is any breach of the terms of the restricted management information.*

- If information is presented in meetings, NSS staff should reiterate verbally that it is for management information purposes only. In such cases, NSS staff should be aware of the attendees and the organisations/groups they are representing, prior to sharing the information.

To help NSS staff determine whether a request for potentially disclosive data can be released, there is a list of considerations contained within the risk assessment – these apply to all types of information to be released, including that intended for management purposes. It is important to note that this protocol does not set a particular formula that provides a measure of risk for every scenario. Rather, the emphasis is on the need for judgement to be made, on a case-by-case basis, of the risk which also applies to all types of information to be released, including that intended for management purposes. There are several key points that should specifically be considered when handling a request for management information:

- is the data output considered sensitive or non-sensitive?
- are there small numbers in the output which are potentially disclosive?
- is an organisation requesting their own data or that of another organisation? (if that of another organisation then the requesting organisation should be able to demonstrate that they have the other’s approval)
- what is the intended audience or distribution e.g. restricted internal, likely to be widely circulated or presented?

### 7.1.1 Potentially disclosive and non-sensitive

If the management information request is for data that is potentially disclosive (i.e. small numbers) and involving sensitive outputs and the intended audience is restricted to the recipient’s own organisation then this is unlikely to pose a high risk of disclosure. The data should be labelled as for management information only (described above) and the standard text attached.
Where it is possible that the audience is not restricted to the recipient’s own organisation then in addition to labelling the data as for management information only and attaching the standard text further steps are needed to reduce the risks of disclosure as much as possible. These might include discussion with the recipient to ensure they are aware of the potential risk of disclosure and the issues surrounding that risk and drawing attention to the standard text to make the recipient fully aware of their responsibility.

In both instances, you must follow the Statistical Disclosure Flow Chart, and document you’re your risk score and decisions in the Disclosure Risk assessment Form.

### 7.1.2 Potentially disclosive and sensitive

If the request is for data that is potentially disclosive and sensitive but the intended audience is restricted to a recipient’s own organisation then this is unlikely to pose a high risk of disclosure. The data should be labelled as management information only and the standard text attached. Steps should be taken to reduce the risks of disclosure as much as possible. These could include discussion with the recipient to ensure they are aware of the potential risk of disclosure and drawing attention to the standard text to make the recipient fully aware of their responsibility. All decisions must be documented in the Disclosure Risk Assessment Form.

Where it is not known if the audience is restricted to the recipient’s own organisation a risk assessment should be completed and a disclosure form submitted. Depending on the outcome of this assessment disclosure control may need to be applied. In these instances, discussion with the user is required to determine whether requirements can be changed or onward distribution controlled while still meeting the user’s needs. The data should be labelled as for management information only and the standard text attached.

Sensitive data to be released to a wide or undefined audience across a number of organisations may pose a high risk of disclosure even if the intended purpose is for management. In some cases, where this information is needed unadjusted, it may be necessary to treat the request in the same way as requests for confidential data (e.g. to request that they complete a Data Request and Linkage Form).

### 7.2 Information released for Data Quality Assurance (QA) purposes (including Peer Review)

Quality assurance, in this instance, refers to the sharing of data with others other than the original data provider, though it may also include them. This refers to outputs, e.g. publications and reports, and not the data collection and maintenance process. A common example is the QA of figures as part of the preparation for public release. When data is shared for quality assurance purposes it may involve organisations that are not the original data provider. If the information is for quality assurance purposes and the data is potentially disclosive then the guidelines provided above which apply to information released for management purposes should also be followed when releasing information for QA purposes.
The label and standard text attached to the data should also be amended for quality assurance purposes as follows:

- Each table within the release should be labelled as ‘Data Quality Assurance only, not for onward distribution’. This statement should appear prominently at the top of each table.
- The standard text below should be attached as a footnote to each table:
  
  *This information has been released for data quality assurance purposes only. The data have not been adjusted to protect against potential disclosure risks and may contain information which enables (perhaps with the aid of further knowledge of the topic) an individual patient or member of staff to be identified. Please ensure circulation is restricted and that data confidentiality is not compromised. Please contact phs.statsgov@nhs.net if you have any queries regarding this or wish to report a breach of the terms of the release.*

- If information is presented in meetings, PHS staff should reiterate verbally that it is for data quality assurance purposes only. In such cases, PHS staff should be aware of the attendees and the organisations/groups they are representing, prior to sharing the information.

### 7.2.1 Information released to a researcher as an intermediate output for the purpose of discussion with colleagues and not for publication

Researchers carrying out data analysis in a safe haven using potentially disclosive data may require the release of intermediate outputs during the course of the project. Access to a safe haven is strictly limited and some research colleagues will have no access. Intermediate outputs are used in discussion with supervisors and colleagues within a research group to make decisions about the direction of a research project. This is particularly useful when the researcher is a PhD student. It is expected that the researcher must have sought permission from the Public Benefit and Privacy Panel which they should evidence. If an output (which could, for example, be a table, a graph or regression output from a statistical package) is cleared for release for this purpose then standard text equivalent to that used above for QA purposes will be attached as a footnote (see below). In particular, it must be made clear to the researcher that the output is not for publication. If an intermediate output is required at the end of the project for a publication, then it must go through the disclosure assessment process again because more stringent criteria may be applied. This would be the case if other outputs requested at the end of the project might, in combination with this one, result in disclosure.

The standard text below should be attached as a footnote to all intermediate outputs released to a researcher:

*This information has been released for internal use by the research team only. The data have not been adjusted to protect against potential disclosure risks and may contain information which enables (perhaps with the aid of further knowledge of the topic) an individual patient or member of staff to be identified. Please ensure circulation is restricted to the named individuals in the research team (as listed in the data access application form) and that data confidentiality is not compromised. This information must not be disseminated beyond the research team and is not for publication.*
8. Annex B - Geographies and Populations

Smaller sized geographies or populations will commonly increase the likelihood of disclosure of information about an individual or group of individuals, and this should be taken into account when assessing the risk of disclosure for any analysis. In addition to the size of the reporting geography or population, PHS staff also need to be aware of residual disclosure or differencing.

Disclosure by differencing occurs whereby the comparison of two or more tables reveals information that is not available in any single table. Another form of disclosure by differencing is geographical differencing, which may be an issue when the boundaries of two geographic areas overlap.

For example, there used to be overlap between NHS board (1st April 2006 configuration) and local authority boundaries. Following the NHS board boundary changes on 1st April 2014 this is no longer an issue. However, the following example highlights the issue if the 2006 NHS board configuration is used for analyses.

If one table contains data relating to a Local Authority and another contains data relating to a Health Board (2006), it is possible that table values may be subtracted to reveal confidential information about the overlapping geography. This is an important issue for PHS to bear in mind as information is frequently published from the same source at differing geographical levels.

There are many different geographical classifications available for analysis in NSS - from postcode level up to Scotland level. It should be noted that not all these geographical classifications are coterminous (i.e. share the same boundaries) and hence gaps or overlaps can occur to create “slivers” or differencing issues. These slivers will potentially have small populations and so care should be taken when analysing and publishing data at varying geographical levels to reduce the risk of any disclosure.

Examples of recent changes regarding geographies:

**NHS Board Boundary changes on 1st April 2014**

NHS Board boundaries changed on 1st April 2014 and now align with those of local authorities. As the boundaries are now aligned, this eliminates the ability to identify small numbers by differencing that previously existed between local authority and 2006 NHS Board boundaries.

However, due to this boundary change consideration must now be given to differencing between NHS Board level statistics previously produced based on the 2006 NHS Board boundaries compared with the same statistics now being produced based on the 2014 NHS Board boundaries.

If differencing in this way causes a disclosure risk, then please contact the geography team for further advice.
Data Zone redraw on 6th November 2014

Data Zones were first created in 2004 and were based on the 2001 Census. On 6th November 2014, Data Zone boundaries were redrawn by the Scottish Government to take changes in population since the first edition into account and ensure a more consistent population size. The new redrawn Data Zones have been based on the 2011 Census, and are known as ‘2011 Data Zones’, with the original Data Zones being known as ‘2001 Data Zones’.

As a result of this, consideration must be given to Data Zone level statistics. Differencing can occur when comparing statistics for a 2001 Data Zone and its closest equivalent 2011 Data Zone. It is possible to work out the difference between the two and assign this to the small non-overlapping area between the two sets of boundaries.

The geography team can be contacted via phs.statsgov@nhs.net
9. Annex C – Examples of the Two Key SDC Methods

9.1 Table Redesign

Example A illustrates the process of table redesign. The first table shows information about the number of people in a local authority who are suffering from illnesses A, B and C by age group. We shall assume that, because the output is sensitive, the data owner considers cell values of less than 5 to be disclosive. There are five such cells in the table, shown in boxes.

In order to protect the table without actually altering the data, the age groups could be combined to form 10-year intervals instead of 5-year intervals. It can be seen that changing the spanning variables in this way has protected the sensitive data and produced a table which can safely be released into the public domain. It is important to be consistent in groupings within variables, between tables produced, to avoid disclosure by differencing.

Example A

<table>
<thead>
<tr>
<th></th>
<th>20-24</th>
<th>25-29</th>
<th>30-34</th>
<th>35-39</th>
<th>40-44</th>
<th>45-49</th>
<th>50-54</th>
<th>55-59</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>5</td>
<td>9</td>
<td>14</td>
<td>11</td>
<td>20</td>
<td>3</td>
<td>14</td>
<td>10</td>
<td>86</td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>24</td>
<td>28</td>
<td>19</td>
<td>10</td>
<td>22</td>
<td>35</td>
<td>39</td>
<td>28</td>
<td>222</td>
</tr>
</tbody>
</table>

9.2 Suppression

The table in Example B contains information about peoples’ ethnicity in Local Authority X, according to their occupation. Assume that the data owner recommends that any cell value of 5 or less should be suppressed. The table shows that one nurse in X is ethnicity B and four teachers are ethnicity D - these cells are disclosive and must therefore be suppressed. Additional suppressions are also required because both row and column totals are shown in the table and can easily be used to work out the missing values.

Example B

<table>
<thead>
<tr>
<th></th>
<th>50&lt;60</th>
<th>60&lt;70</th>
<th>70&lt;80</th>
<th>80&lt;90</th>
<th>90&lt;100</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lawyer</td>
<td>16</td>
<td>7</td>
<td>11</td>
<td>21</td>
<td>6</td>
<td>61</td>
</tr>
<tr>
<td>Nurse</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>60</td>
</tr>
<tr>
<td>Police</td>
<td>11</td>
<td>13</td>
<td>19</td>
<td>9</td>
<td>15</td>
<td>67</td>
</tr>
<tr>
<td>Teacher</td>
<td>9</td>
<td>8</td>
<td>12</td>
<td>4</td>
<td>14</td>
<td>47</td>
</tr>
<tr>
<td>Total</td>
<td>46</td>
<td>29</td>
<td>48</td>
<td>51</td>
<td>51</td>
<td>225</td>
</tr>
</tbody>
</table>
If we first consider the single nurse earning between £60k and £70k, it can be seen that two secondary suppressions are required to protect this cell - one from the occupation row and another from the income column. The same is also true of the disclosive ‘teacher’ cell. However, due to the positioning of these two disclosive cells, it is in fact possible to protect them simultaneously using only two secondary suppressions instead of four, thus preserving more of the original data.

<table>
<thead>
<tr>
<th></th>
<th>50&lt;80</th>
<th>60&lt;70</th>
<th>70&lt;80</th>
<th>80&lt;90</th>
<th>90&lt;100</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lawyer</td>
<td>16</td>
<td>7</td>
<td>11</td>
<td>21</td>
<td>6</td>
<td>61</td>
</tr>
<tr>
<td>Nurse</td>
<td>10</td>
<td>*</td>
<td>6</td>
<td>*</td>
<td>16</td>
<td>50</td>
</tr>
<tr>
<td>Police</td>
<td>11</td>
<td>13</td>
<td>19</td>
<td>9</td>
<td>15</td>
<td>67</td>
</tr>
<tr>
<td>Teacher</td>
<td>9</td>
<td>*</td>
<td>12</td>
<td>*</td>
<td>14</td>
<td>47</td>
</tr>
<tr>
<td>Total</td>
<td>46</td>
<td>29</td>
<td>48</td>
<td>51</td>
<td>51</td>
<td>225</td>
</tr>
</tbody>
</table>

The second table shows how the secondary suppressions have successfully protected the data by removing two cell values, thus preventing disclosure by subtraction.

In practice, there are usually more than two disclosive cells, which can cause the number of secondary suppressions required to rise substantially.
10. Annex D – Statistical Disclosure Control Flowchart

Do any of these conditions apply?

- Counts from 1 to 9 exist
- Columns and rows dominated by zeros or 100% rates
- Population or geography classed as small (e.g. island Boards, hospital level, GP Practices, clinics)
- An individual (or groups) can be identified OR additional personal or sensitive information about an individual (or groups) can be gained from the data (e.g. age, postcode, sexual orientation, ethnicity of an individual, data about children broken down by age group and NHS Board, and grounds for abortion)

No

Complete disclosure risk assessment

Risk Assessment Score < 4

Risk Assessment Score ≥ 4

Yes

Do the outputs reveal sensitive information about an individual?

No

Release

Yes

Protect

DISCLOSURE RISK ASSESSMENT

For likelihood of an attempt to disclose, consider:
- Who wants the data?
  - How will it be accessed?
- What data do they want?
- Why do they want the data?
- Similar known data out there?

For impact of disclosure, consider:
- Is the output sensitive?
- What data do they want?
- Will additional personal or sensitive information be gained about an individual?

Complete the Disclosure Risk Assessment Form and retain it as evidence in your shared electronic folders.

CELL PROTECTION METHODS

Table redesign - increase age or other grouping, group Island Boards, aggregate tables across a number of years, months or quarters, etc.

Cell suppression – For primary suppression, replace values 1 to 4 with asterisk (include values up to 9, if required, for any combination of small population (geography and high sensitivity and columns dominated by zeros). Use secondary suppression to suppress next higher cell value on column if primary suppressed cell values can be derived by subtraction. Use the same symbol for both primary and secondary suppressed cells. For open data, use ‘c’ instead of asterisk.

For open data or other suppression methods, i.e. rounding and cell swapping, contact Stats Governance Team for guidance at: phs.statsgov@nhs.net

SENSITIVE OUTPUTS

Outputs about a sensitive topic may have certain characteristics which are not sensitive (e.g. tables showing waiting times for drug misuse treatment).

Sensitive topics and populations include but are not limited to:
- Sexually transmitted disease
- Abortion
- Suicide, self-harm
- Drug and alcohol misuse
- Mental health
- Contraceptive prescriptions
- Vulnerable populations e.g. people subject to Looked After Children (Scotland) Regulations 2009 and Adult Support & Protection (Scotland) Act 2007

This form should be completed for all statistical releases and information requests where the Disclosure Flowchart has instructed to undertake a Risk Assessment and/or Protect. The completed form, once signed off, should be stored locally for further reference. Please use the Statistical Disclosure Control Protocol for guidance on completing the form.

Publication Title, Project or Information Request Ref. No: ________________________________

Name: ___________________________________________  Date: ___________________________
Team: ___________________________________________

PART ONE: RISK ASSESSMENT - only complete as instructed on Flowchart

Indicate below what you think the likelihood of disclosure is and the level of impact that this disclosure would cause:

<table>
<thead>
<tr>
<th>Likelihood of an attempt to disclose</th>
<th>Impact of disclosure</th>
</tr>
</thead>
<tbody>
<tr>
<td>□ High (3)</td>
<td>□ High (3)</td>
</tr>
<tr>
<td>□ Medium (2)</td>
<td>□ Medium (2)</td>
</tr>
<tr>
<td>□ Low (1)</td>
<td>□ Low (1)</td>
</tr>
</tbody>
</table>

Questions to consider when undertaking a risk assessment include:

- Exactly what data is requested? Data sources, variables, time periods etc
  - a. How sensitive is the topic area considered? For data initially deemed non-sensitive, if small numbers are present then the topic sensitivity should be given careful thought. Could the information be considered sensitive by others?
  - b. What are the sizes of the geographies / populations / institutions involved?
  - c. Consider the size of cell values / is the table design most appropriate? See protocol for guidance / also take guidance from previous publications or requests.

- Who is requesting the data? Named contact, position/role, organisation

- What is the intended use of the data? e.g. inform committee meeting, FOI, PQ, research paper, publication, SG policy.

- Who will have access to the data? Named contact(s)/groups/organisation, position/role, NHS board, NHS Steering Group, Scottish Government policy makers, non-NHS partner organisations
  - a. Will the data be in the public domain? e.g. Information request to media, PQ, FOI
  - b. Will there be controlled access?

- What measures are there in place to protect the information? e.g. none (info will be in public domain); info will be distributed at meeting only; info will be distributed within SG only.

If by multiplying the likelihood by impact the score is 4 or above, disclosure control methods should always be applied prior to release. If the risk score is less than 4, then the data can be released however you should complete the relevant questions in Part Two of this form. In some circumstances, you may also wish to consider applying disclosure control for lower scores especially if the topic is sensitive.

Now complete part two of this form.
PART TWO: DETAILS OF DISCLOSURE CONTROL - complete for all statistical releases and IRs where a Risk Assessment has been undertaken OR protection of information is instruct as indicated on the Flowchart.

1. Did you apply any Disclosure Control techniques prior to releasing this data?
   - □ Yes, go to question 2
   - □ No, go to question 5 to explain why no disclosure control was applied

2. Indicate what the potential disclosure risks were with this data (tick all that apply):
   - □ Cells with values of 1 - 4
   - □ Cells with values of 5 - 9
   - □ Sensitive output
   - □ Identify individuals & gain additional personal or sensitive information
   - □ Rows/Columns dominated by zeros
   - □ 1 or 2 hospitals/practitioners
   - □ Small population/geography or institution level
   - □ Other, please specify: ____________________________________________

3. Indicate which disclosure control techniques you applied:
   - □ Table Redesign
   - □ Cell suppression
   - □ Other, please specify (Head of Statistical Governance Team must approve other methods):

4. Explain below the impact of applying disclosure control on the various tables of data. If this is different for each table, please document details for each. You should also explain why the method(s) were chosen.

5. Additional comments (please use the space below to provide any further relevant information. Use this space to document why you did not decide to apply any disclosure control prior to release.)

6. Sign off by service manager, information consultant, team manager, or equivalent, or their authorised deputy (no need to submit this form; keep it for your records):

   Name: ___________________________       Date: ___________________________