Office for National Statistics

SRS Output Checking Guidance Document

Work strand: Statistical Disclosure Control

Release: v.1.1

Date: 05.09.2022

Owner: Head of Statistical Support

Document ID:

Version No: v.1.0

Document History

Revision History

Date of this revision: 05.09.2022

Date of next revision: 05.03.2023

Revision date	Summary of Changes	Changes marked
05.09.2022	First issue	
12.06.2023	First revision	

For queries relating to this document please contact the Statistical Support Team at statistical.support@ons.gov.uk.

Office for National Statistics

Table of Contents

1. Clearance types 5
2. General output guidance 5
3. ‘Safe’ and ‘unsafe’ outputs 6
4. Default SDC ‘rules of thumb’ 7
5. File types 8
6. Frequency tables and other tables 10
6.1 Low counts and zeros 10
6.1.1 Example: suppression 10
6.1.2 Example: rounding 11
6.1.3 Example: reformatting 12
6.2 Class disclosure 13
6.2.1 Example: structural zeros 13
6.2.2 Example: suppression 14
6.2.3 Example: rounding 14
6.2.4 Example: reformatting 15
6.3 Secondary disclosure 15
6.3.1 Example: re-calculating totals 16
6.3.2 Example: secondary suppression 17
6.3.3 Example: rounding 18
7. Dominance 18
7.1 Example: reformatting 19
8. Statistics 20
8.1 ‘Safe’ statistics 20
8.2 ‘Unsafe’ statistics 20
8.2.1 Mean 20
8.2.1.1 Example: suppression 21
8.2.2 Percentages 22
8.2.3 Weighted counts 22
8.2.4 Mode, minimums and maximums 22
8.2.4.1 Example: suppression 23
8.2.4.2 Example: rounding 24
8.2.5 Medians, quartiles, deciles and percentiles 24
8.2.5.1 Example: suppression 25

8.2.5.2 Example: reformatting 26
8.2.6 Ratios, including odds ratios 26
8.2.6.1 Example: suppression 26
9. Graphs 27
9.1 Line graphs 27
9.1.1 Example: suppression 27
9.1.2 Example: reformatting 29
9.2 Scatter graphs 29
9.2.1 Example: reformatting 30
9.3 Bar charts and histograms 31
9.3.1 Example: reformatting 31
9.3.2 Example: suppression 33
9.4 Boxplots 33
9.4.1 Example: suppression of plots 34
9.4.2 Example: reformatting plots 35
9.4.3 Example: suppression of outliers 36
9.4.4 Example: reformatting whiskers 36
9.5 Violin plots 37
9.5.1 Example: reformatting plots 37
10. Regressions and modelling 39
10.1 Coefficients, margin plots and test statistics 39
10.1.1 Example: saturated regression 39
10.2 Residuals 40
11. Maps and spatial analysis 40
11.1 Maps 40
11.1.1 Example: reformatting 41
11.2 Geographies 42
12. Code files 42
12.1 Example: Hard-coded data in code 43
12.2 Example: Disclosive comments in code 43
12.3 Example: Data table in code 44
12.4 Example: Overly specific code 44

Office for National Statistics

1. Clearance types

Within the SRS, there are three types of clearance offered. Researchers must clearly state, within their Output Request form **and** within their requesting email, which type of clearance they want for their output. The three types are as follows:

- **Pre-Publication ('PrePub') clearance:** suitable for any file type. The files may **only** be shared with researchers, sponsors and/or funders who are **named** on the project; the files must be deleted once the project ends.
- **Publication ('Pub') clearance:** suitable for any file type that is publication ready. The files may be shared beyond individuals named on the project; the files may be retained indefinitely.
- **Code clearance:** only suitable for code files that do not contain **any** SRS data. The files may be shared beyond individuals named on the project; the files may be retained indefinitely; the disclaimer is less extensive than for Pub clearances.

In all cases, when the cleared output is sent to the researcher, the email will contain the appropriate disclaimer. On receipt of the file(s), the researcher **must** add this disclaimer text to them.

2. General output guidance

The SRS operates using principles-based output Statistical Disclosure Control (PBOSDC). This means that there are several SDC 'rules of thumb' which are followed as default – see section 4 for an overview and sections 6 to 12 for more detailed guidance. However, in principle **any** output is allowed provided it does not result in meaningful disclosure.

The researcher may request an exemption to the SDC 'rules of thumb' for **any** output. Requests for exemptions will be assessed by the Statistical Support staff on the basis of whether:

- The output is highly important (it is required to enable the project to provide its planned research for the public good) **and**
- Any deviations from the SDC 'rules of thumb' do not result in **meaningful** disclosure beyond that permitted by the Data Owner for the given project.

Please note that if an exemption is requested, **all** necessary information for allowing the 'rules of thumb' (e.g., underlying unweighted counts) **must** be provided up front, as it would be if an exemption was not being requested. A statement informing output checkers that you are requesting an exemption to the SDC 'rules of thumb' and a justification for the exemption **must** also be provided, with the latter covering **both** of the two points above. This should be provided in the description section at the end of the Output Request form.

Operating the SRS using a PBOSDC system is only possible if researchers restrict their requests for exemptions to the SDC 'rules of thumb' to **very** occasional instances involving **highly** important outputs.

Following the PBOSDC system, there will very occasionally be instances where the particular context and content of the output means that the output will require **stricter** SDC than the SDC 'rules of thumb'. If this occurs, the output checkers will clearly outline the disclosure risk that has been identified and will assist the researcher to apply stricter SDC to appropriately manage this disclosure risk.

3. 'Safe' and 'unsafe' outputs

The guidance in the following sections has been written using the concept of 'safe' and 'unsafe' outputs in SDC literature. This classification system is used to ensure that researchers' time applying SDC and output checkers' time checking SDC is focused on the outputs which are most likely to represent a disclosure risk. Under this system, the categories are defined as follows[1]:

- 'Safe' outputs: **will** be released **unless** SDC checks demonstrate some reason why they should be held back or adjusted – the SDC 'rules of thumb' outlined below are fairly minimal as they are designed to enable checks for these instances of potential risk.
  The researcher should always provide the minimum information required (e.g., total count of data subjects for regressions and models) but can generally expect the output to be cleared with minimal or no further changes. I.e., the burden of proof is on the output checker to provide reason(s) why the output **cannot** be released, contrary to normal expectations for this type of output.
- 'Unsafe' outputs: **will not** be released **unless** the researcher can demonstrate, via appropriate SDC and, where applicable, contextualising information, that the output meets the detailed criteria for this type of output – the SDC 'rules of thumb' outlined below represent these detailed criteria.
  The researcher should always provide the minimum information required. However, the output will not be cleared unless the researcher demonstrates, to the output checkers' satisfaction, that the particular context and content of the output makes it non-disclosive. I.e., the burden of proof is on the researcher to provide reason(s) why the output **can** be released – generally, **but not always**, appropriate contextualising information (e.g., clear variable labels, graph titles, etc.) and the SDC 'rules of thumb' will ensure that sufficient reasons are provided.

In general, statistics are classified as 'safe' or 'unsafe' as follows[2]:

Type of statistics	Type of Output	Classification
Descriptive statistics	Frequency tables	'Unsafe'
	Magnitude tables	'Unsafe'
	Means	'Unsafe'
	Percentages	'Unsafe'
	Weighted counts	'Unsafe'
	Mode, minima and maxima	'Unsafe'
	Medians, quartiles, percentiles	'Unsafe'
	Indices, ratios, indicators	'Unsafe'
	Concentration ratios (including Herfindahl-Hirschman Index (HHI))	'Safe'

[1] This definition is adapted from:
Ritchie F. (2008) "Disclosure detection in research environments in practice", in Work session on statistical data confidentiality 2007; Eurostat; pp. 399-406.
Brandt M. *et al*. (2010) "Guidelines for the checking of output based on microdata research", Final report of ESSnet subgroup on output SDC.

[2] This table is adapted from: Brandt M. *et al*. (2010) "Guidelines for the checking of output based on microdata research", Final report of ESSnet subgroup on output SDC.

	Higher moments of distributions (including variance, covariance, kurtosis and skewness)	'Safe'
	Graphs: pictorial representations of actual data	'Unsafe'
Correlation and regression analysis	Linear regression coefficients	'Safe'
	Non-linear regression coefficients	'Safe'
	Estimation residuals	'Unsafe'
	Summary and test statistics from estimates (R^2, χ^2, etc.)	'Safe'
	Correlation coefficients	'Safe'

4. Default SDC 'rules of thumb'

As described in section 2, the SRS operates using PBOSDC and therefore has SDC 'rules of thumb' rather than absolute SDC rules. These apply in all cases except where:

a) There are dataset-specific SDC rule(s) set by the Data Owner (see the spreadsheet within the SRS at Libraries$/SRS and SDC Guidance/SDC Guidance by Dataset for a full list of SDC rules by dataset). In these instances, the dataset-specific SDC rule(s) would be used instead, superseding the SRS's SDC 'rules of thumb'.
b) An exemption has been granted permitting the project to have custom SDC rules. In these instances, the project-specific SDC rule(s) would be used instead, superseding both the SRS's SDC 'rules of thumb' and the dataset-specific SDC rule(s).
c) An exemption is requested as described in section 2.

The default SDC 'rules of thumb' for the SRS are:

1) The output must be accompanied by a completed Output Request form.
2) The output (files to be cleared, Output Request form and any supplementary files) must be sufficiently clear and comprehensible to permit SDC checking without the need for dataset- or project-specific knowledge. E.g., variable names should be self-explanatory or explained, tables and figures should be appropriately labelled, non-SRS data should be clearly labelled, etc.
3) The threshold is 10 data subjects. Unweighted counts <10 and statistics (e.g., weighted counts, percentages, means, etc.) derived from groups containing <10 data subjects must be suppressed. Zeros are included in this threshold – i.e., counts ranging from zero to nine, inclusive, are not permitted.
    a) In this context, 'data subject' refers to individuals (e.g., employees, pupils, teachers, patients) and/or organisations (e.g., firms, schools, nurseries, universities, charities). Counts must be provided at **data subject level**. Providing the number of geographical areas (e.g., wards), aggregate counts (e.g., industries) or time period(s) (e.g., number of days) the statistic relates to would not be sufficient.
    b) As there is a threshold, by definition record-level data is not permitted in outputs.
4) Class disclosure is not permitted.
5) Dominance is not permitted.
6) There must not be any way to reverse SDC, e.g., by differencing. If this is possible, secondary SDC, e.g., secondary suppression or reformatting, must be applied to prevent it.

If the researcher does not wish to include this information within their output file, particularly underlying unweighted counts, it may be provided in supplementary file(s) that are used for SDC checks but not cleared from the SRS. These file(s) should be clearly titled. A note should be placed in the description section of the Output Request form explaining which file(s) are for clearance and which are not, plus which file(s) contain the underlying counts for which output file(s).

If complex rules are used, particularly custom rules involving specific rounding rules, SDC checks often run smoothest when an additional copy of the output file(s) are provided for reference, showing the data before SDC was applied. If this is done, please add an explanatory note in the description section of the Output Request form to avoid confusion.

If an output contains any data that is non-SRS data, this must be clearly labelled – within the file and/or via a note in the in the description section of the Output Request form. Non-SRS data does not have to conform to the SRS's SDC 'rules of thumb'. However, depending on the circumstances, written confirmation of Data Owner approval for the output may be required.

5. File types

We have the capacity to check a wide variety of file types. These include:

- Data files, such as:
  - .csv, .tsv (generic)
  - .xls, .xlsx (Microsoft Excel)
  - .sav (SPSS)
  - .dta (STATA)
  - .sd2, .sas7bdat, .sd7 (SAS)
  - .Rdata, .Rda, .Rds (R) – when they contain a single dataframe/table
- Log files, such as:
  - .log (generic)
  - .smcl (STATA)
  - .spo, .spv (SPSS)
- Image files, such as:
  - .bmp, .gif, .jpeg, .png (generic)
  - .gph (STATA)
- Document and presentation files, such as:
  - .doc, .docx (Microsoft Word)
  - .ppt, .pptx (Microsoft PowerPoint)
- Code files, such as:
  - .do (STATA)
  - .sps (SPSS)
  - .sas (SAS)
  - .R (R)
  - .py, .py3 (Python)
  - .ipynb (Jupyter Notebook)
- Files associated with coding packages, such as:

  - .ado, .mata, .pkg, .sthlp, .toc (STATA)
  - .tar.gz, .tar (R)
- Miscellaneous other files, including:
  - .txt
  - .pdf
  - .TeX

There are some file types which we do not have the capacity to fully check. These file types **may not be cleared for any outputs**, regardless of their apparent content:

- Markdown files – with .html extension if in Hypertext Markup Language, .markdown, .md, .markdn or .mdown extensions if in Markdown language, with .Rmd extension if in R or with .Rnw extension if in LaTeX.
  - Used to format the structure of webpages, for writing code documentation and for dynamic reports, documents, presentations, dashboards, websites, etc. Combine code, documentation, data and/or metadata into a single file.
  - As each file is unique, due to the wide range of styles and types of information that can be held within them, we cannot guarantee that we can exhaustively check these files.
  - Instead, the researcher should provide the information in one of the above checkable formats – they may reconstitute them back into a markdown file once outside the SRS, if they wish.
- R project files – with .Rproj extension.
  - Saves an entire project, including data files, code and outputs.
  - Used to store a project neatly in one place, improving workflow.
  - As each project will contain different things, potentially including data and outputs, we cannot guarantee that we can exhaustively check these files.
  - Instead, the researcher should provide the information in constituent .R, .Rdata, .Rda and/or .Rds files as these are checkable – they may reconstitute them back into an .Rproj file once outside the SRS, if they wish.
- Shapefiles – with .shp, .shx or .dbf extensions or, more occasionally, with .prj, .sbn, .sbx, .fbn, .fbx, .ain, .aih, .ixs, .mxs, .atx, .shp.xml, .cpg or .qix extensions.
  - Digital vector storage formats and associated supporting files.
  - Predominantly used for storing geographic data.
  - We can open these types of files, but their structure makes them exceptionally difficult to check, so we cannot guarantee that we can exhaustively check these files.
  - Instead, the researcher should provide the data in one of the above checkable formats – they may reformat this back into a shapefile once outside the SRS, if they wish.
- JavaScript Object Notation files – with .json extension.
  - A language-independent data storage file, in which the data is organised as a hierarchical list.
  - We can open this type of file, but its structure makes it exceptionally difficult to check, so we cannot guarantee that we can exhaustively check these files.
  - Instead, the researcher should provide the data in one of the above checkable formats – they may reformat this back into a .json file once outside the SRS, if they wish.

- R presentation files – with .RPres extension.
    - Uses Markdown and R code to create HTML5 presentations.
    - Like other files using Markdown, e.g. .Rmd files, we cannot guarantee that we can exhaustively check these files.
    - Instead, the researcher should provide the data in one of the above checkable formats – they may reformat this back into an .RPres file once outside the SRS, if they wish.
- R data files – with .Rdata, .Rda, .Rds extensions – that contain data in formats other than dataframes/tables:
    - A data storage file for R coding language.
    - We can open this type of file, but unless it contains a single dataframe/table the structure is generally too complex and not sufficiently human-readable. Therefore, we cannot guarantee that we can exhaustively check these files when they contain data in formats other than a single dataframe/table.
    - Instead, the researcher should provide the data in one of the above checkable formats.

6. Frequency tables and other tables

6.1 Low counts and zeros

Any unweighted counts must meet or exceed the threshold. Any statistics must relate to a group whose unweighted count meets or exceeds the threshold – this includes percentages and weighted counts – see section 8. In order to assess this, these unweighted counts must be clearly provided to demonstrate that appropriate SDC has been applied. Note that 'group' refers to the numerator, not the denominator when information is reported about categories. E.g., if the mean age was reported for males and females in a study, the count of the 'group' would refer to the number of males and the number of females, not the total number of study participants.

Zeros are considered a disclosure risk, unless it is evident from the output request that they are structural – see section 6.2.

Any unweighted counts below the threshold must be suppressed or otherwise removed, e.g., by reformatting the table. Likewise, statistics relating to a group whose unweighted count is below the threshold must also be suppressed. If any values are suppressed, secondary SDC must be applied to prevent secondary disclosure of the suppressed values by differencing – see section 6.3.

6.1.1 Example: suppression

This table contains numerous instances of low counts (highlighted in yellow to make them clearer). SDC should be applied to this table.

Medical condition	Jul-20	Aug-20	Sep-20	Oct-20	Nov-20	Dec-20	Total
None	12	19	22	10	9	11	83
Prefer not to say	32	42	37	31	29	24	195
Allergy	21	15	24	17	13	6	96
Viral	8	19	11	11	14	16	79
Bacterial	10	9	23	17	13	12	84
Cancer	8	3	12	9	18	1	51
Arthritis	1	5	2	6	2	5	21
Hereditary condition	18	1	8	13	10	3	53
Total	110	113	139	114	108	78	662

If suppression is chosen as the SDC method, this would result in the following table:

Medical condition	Jul-20	Aug-20	Sep-20	Oct-20	Nov-20	Dec-20	Total
None	12	19	22	10	-	11	83
Prefer not to say	32	42	37	31	29	24	195
Allergy	21	15	24	17	13	-	96
Viral	-	19	11	11	14	16	79
Bacterial	10	-	23	17	13	12	84
Cancer	-	-	12	-	18	-	51
Arthritis	-	-	-	-	-	-	21
Hereditary condition	18	-	-	13	10	-	53
Total	110	113	139	114	108	78	662

'-' indicates suppression due to low counts.

Here '-' has been chosen by the researcher to indicate suppression. Other common ways include '*', '.', 'x', '<10' and 'SUPP'. The exact symbols or letters chosen to indicate suppression is up to the researcher. The only proviso is that they must not enable the reader to crack the suppression – e.g., suppressing counts of 1-9 with '<10' but zeros with '-' would not be acceptable. Some Data Owners might have specific rules for which symbols to use, therefore please consider this when applying suppression to your outputs.

In this instance, this initial (primary) suppression is not sufficient on its own to prevent disclosure. The totals highlighted in yellow, in combination with the other values in those rows, enable recalculation of the suppressed values in that row. For example, in the 'None' row, the total of 83 minus the other counts (12, 19, 22, 10 and 11) informs us that the suppressed value is nine. This is termed **secondary disclosure**. To prevent this, secondary SDC should be applied – see section 6.3.

6.1.2 Example: rounding

Alternatively, the researcher could choose rounding as the SDC method, which would result in the following table:

Medical condition	Jul-20	Aug-20	Sep-20	Oct-20	Nov-20	Dec-20	Total
None	10	20	20	10	10	10	80

Prefer not to say	30	40	40	30	30	20	200
Allergy	20	20	20	20	10	10	100
Viral	10	20	10	10	10	20	80
Bacterial	10	10	20	20	10	10	80
Cancer	10	0	10	10	20	0	50
Arthritis	0	10	0	10	0	10	20
Hereditary condition	20	0	10	10	10	0	50
Total	110	110	140	110	110	80	660

All values have been rounded to the nearest 10. Counts may not sum to totals due to rounding.

The rounding rule(s) are chosen by the researcher (unless specified under dataset- or project-specific SDC rules). In this case, the researcher has chosen to round to the nearest 10. Here, the zeros are permitted as they do not exclusively represent **real** zero counts, but rather any count from 0-4 (with counts of 5-9 rounded up to 10, as indicated by the annotation below the table).

Variations on rounding are permitted – e.g., rounding all of the raw counts but not the totals or only rounding the counts that are below threshold. However, it must be made clear to the output checker and reader how the rounding has been applied, so that it is clear which counts are unrounded and which are rounded.

6.1.3 Example: reformatting

A third option available to the researcher is to reformat the table, which could result in the following table:

Medical condition	Jul-20	Aug-20	Sep-20	Oct-20	Nov-20	Dec-20	Total
None / prefer not to say	44	61	59	41	38	35	278
Short-term condition, e.g., viral, bacterial	18	28	34	28	27	28	163
Longer-term condition, e.g., allergy, cancer, arthritis, hereditary condition	48	24	46	45	43	15	221
Total	110	113	139	114	108	78	662

The way that the table is reformatted is chosen by the researcher. Reformatting can include any of the following:

- Merging rows together (as shown in the example above).
- Merging columns together.
- Removing rows and then recalculating totals.
- Removing columns and then recalculating totals.

The aim, as with all SDC, is to preserve as much information as possible. This is why the exact method of reformatting is chosen by the researcher – they are best placed to know how to reformat whilst ensuring that the results or key points of the analysis are retained.

6.2 Class disclosure

Class disclosure occurs when the reader can learn something new about every data subject belonging to a particular group of the data. For this reason, empty cells (i.e., cells whose unweighted counts are zero) and full cells (i.e., cells whose unweighted counts represent 100% of a group) represent a class disclosure risk.

The exception to this is **structural zeros**, also called **logical zeros**. These are zeros which are present due to the nature of the dataset or its collection – i.e., the only possible value of the group is zero – see section 6.2.1 for an example. Structural zeros are permitted.

Unweighted counts of zero, excepting structural zeros, must be suppressed or otherwise removed, e.g., by reformatting the table. Unweighted counts representing 100% of a group and statistics that relate to 100% of a group may be permitted or may need to be suppressed, depending on the amount of information gained from the class disclosure. If any values are suppressed, secondary SDC must be applied to prevent secondary disclosure of the suppressed values by differencing – see section 6.3. If dataset- and/or project-specific knowledge is required to know that a zero is structural, the researcher should annotate their file and/or provide a note in the description section of the Output Request form explaining this, to prevent confusion during SDC checking.

6.2.1 Example: structural zeros

Sometimes, zeros are the only possible value for a group. These are termed structural zeros and may be included in outputs. For example:

Highest qualification	Age (years)				Total
	0-15	16-20	21-25	26+	
Higher education	0	0	165	148	313
Secondary education	0	152	210	318	680
None	324	65	42	15	446

The zeros (highlighted in yellow to make them clearer) make logical sense. You would not expect anyone younger than 16 years old to have gained secondary education qualifications and would not expect anyone younger than 21 years old to have gained a higher education qualification. The zeros are, therefore, structural rather than informative.

Due to their nature, structural zeros present no meaningful disclosure risk. In contrast, informative zeros (zeros where counts are able to be a value other than zero) do present a disclosure risk. Hence, structural zeros are permitted whereas informative zeros are considered a potential class disclosure risk.

If dataset- and/or project-specific knowledge is required to know that a zero is structural, the researcher should annotate their file and/or provide a note in the description section of the Output Request form explaining this, to prevent confusion during SDC checking.

Non-responses are not necessarily considered structural zeros, context of the output is required and requests will be considered on a case by case basis.

6.2.2 Example: suppression

It is quite common for frequency tables to contain informative zeros. For example:

Highest qualification	Income quartile (lowest to highest)				Total
	1	2	3	4	
Post-graduate	0	0	11	16	27
Degree	10	12	24	27	73
College	11	38	26	10	85
School	43	39	0	0	82
None	54	12	0	0	66
Total	118	101	61	53	333

This table contains several informative zeros (highlighted in yellow to make them clearer). These are disclosive as they enable us to gain specific information about every individual in a group. E.g., we can see that nobody whose highest qualification was school-level or less is in the 3rd or 4th income quartile (i.e., earns above the median) and we can also see that nobody whose highest qualification was post-graduate is in the 1st or 2nd income quartile (i.e., earns below the median). Therefore, SDC should be applied to this table.

If suppression is chosen as the SDC method, this would result in the following table:

Highest qualification	Income quartile (lowest to highest)				Total
	1	2	3	4	
Post-graduate	SUPP	SUPP	11	16	27
Degree	10	12	24	27	73
College	11	38	26	10	85
School	43	39	SUPP	SUPP	82
None	54	12	SUPP	SUPP	66
Total	118	101	61	53	333

'SUPP' indicates suppression due to low counts.

In this instance, the initial (primary) suppression is not sufficient on its own to prevent disclosure. The totals highlighted in yellow, in combination with the other values in those rows, enable recalculation of the suppressed values in that row. For example, in the 'School' row, the total of 82 minus the other counts (43, 39) is zero, which informs us that both of the suppressed values are zero. This is termed **secondary disclosure**. To prevent this, secondary SDC should be applied – see section 6.3.

6.2.3 Example: rounding

Alternatively, the researcher could choose rounding as the SDC method. Rounding may be used as an SDC method for class disclosure, provided that the rounding suitably disguises which zeros are real zeros and which are low counts rounded to zero. SDC by rounding would result in the following table:

Highest qualification	Income quartile (lowest to highest)				Total
	1	2	3	4	
Post-graduate	0	0	10	20	30

Degree	10	10	20	30	70
College	10	40	30	10	90
School	40	40	0	0	80
None	50	10	0	0	70
Total	120	100	60	50	330

All values have been rounded to the nearest 10. Counts may not sum to totals due to rounding.

As discussed in section 6.1.2, the rounding rule(s) are chosen by the researcher (unless specified under dataset- or project-specific SDC rules). Also, variations on rounding are permitted – e.g., rounding all of the raw counts but not the totals or only rounding the counts that are below threshold. However, it must be made clear to the output checker and reader how the rounding has been applied, so that it is clear which counts are unrounded and which are rounded.

6.2.4 Example: reformatting

A third option available to the researcher is to reformat the table, which could result in the following table:

Highest qualification	Income quartile (lowest to highest)				Total
	1	2	3	4	
Degree or above	10	12	35	43	100
College or school	54	77	26	10	167
Total	64	89	61	53	267

The way that the table is reformatted is chosen by the researcher. Reformatting can include any of the following:

- Merging rows together (as shown in the example above).
- Merging columns together.
- Removing rows and then recalculating totals (also shown in the example above).
- Removing columns and then recalculating totals.

The aim, as with all SDC, is to preserve as much information as possible. This is why the exact method of reformatting is chosen by the researcher – as they are best placed to know how to reformat whilst ensuring that the results or key points of the analysis to be retained.

6.3 Secondary disclosure

If values have been suppressed in a table, the table must be checked to ensure that secondary disclosure cannot occur (i.e., the suppressed values cannot be re-calculated, aka differenced, from the remaining values in the output). In particular, the researcher should check for percentages or totals, either within the table or elsewhere in the output file(s), that could enable differencing. However, any statistic that is calculated using a simple formula may result in differencing – e.g., means, weighted counts, ratios, odds ratios.

If differencing can occur, secondary SDC must be applied (i.e., suppression of additional aspects of the output that are not disclosive in and of themselves, but which permit differencing). The researcher should decide how to apply secondary SDC, as in most cases several options are available and the choice between them will depend on which parts of the output are most important for use.

Special care should be taken to ensure that the **whole** output is checked for the possibility of differencing, not just the table with the suppressed values. Secondary disclosure may occur due to a combination of tables or the combination of tables and text which contains numbers. If the scope of differencing in the output is extensive and/or complex, we strongly suggest rounding is used as the secondary SDC technique as this method is the strongest protection against differencing and is easier to comprehensively apply and check.

6.3.1 Example: re-calculating totals

Totals are often the source of differencing and therefore addressing totals is an alternative way of carrying out secondary SDC.

This table was created in section 6.1.1 and contains secondary disclosure. The totals highlighted in yellow, in combination with the other values in those rows, enable recalculation of the suppressed values in that row. For example, in the 'None' row, the total of 83 minus the other counts (12, 19, 22, 10 and 11) informs us that the suppressed value is nine.

Medical condition	Jul-20	Aug-20	Sep-20	Oct-20	Nov-20	Dec-20	Total
None	12	19	22	10	-	11	83
Prefer not to say	32	42	37	31	29	24	195
Allergy	21	15	24	17	13	-	96
Viral	-	19	11	11	14	16	79
Bacterial	10	-	23	17	13	12	84
Cancer	-	-	12	-	18	-	51
Arthritis	-	-	-	-	-	-	21
Hereditary condition	18	-	-	13	10	-	53
Total	110	113	139	114	108	78	662

'-' indicates suppression due to low counts.

Secondary SDC must be applied to this table to prevent the secondary disclosure. If recalculation of totals is chosen as the secondary SDC method, the following table is produced:

Medical condition	Jul-20	Aug-20	Sep-20	Oct-20	Nov-20	Dec-20	Total
None	12	19	22	10	-	11	74
Prefer not to say	32	42	37	31	29	24	195
Allergy	21	15	24	17	13	-	90
Viral	-	19	11	11	14	16	71
Bacterial	10	-	23	17	13	12	75
Cancer	-	-	12	-	18	-	30
Arthritis	-	-	-	-	-	-	-
Hereditary condition	18	-	-	13	10	-	41
Total	93	95	129	99	97	63	576

'-' indicates suppression due to low counts. Totals have been recalculated.

As the totals have been recalculated and now only include the counts that were not suppressed, it is now impossible to recalculate the suppressed counts using them. Note that **all** of the totals have been recalculated, not just those that could be used to recalculate suppressed counts. Having a

consistent approach to how totals are calculated in a given table is important to prevent output checker and reader confusion.

6.3.2 Example: secondary suppression

As an example, we will use the table created in section 6.1.1 again. Another method of secondary SDC is secondary suppression, which could result in the following table:

Medical condition	Jul-20	Aug-20	Sep-20	Oct-20	Nov-20	Dec-20	Total
None	12	19	22	-	-	11	83
Prefer not to say	32	42	37	31	29	24	195
Allergy	21	15	24	17	-	-	96
Viral	-	19	-	11	14	16	79
Bacterial	10	-	23	17	13	-	84
Cancer	-	-	12	-	18	-	51
Arthritis	-	-	-	-	-	-	21
Hereditary condition	18	-	-	13	10	-	53
Total	110	113	139	114	108	78	662

'.' indicates suppression due to low counts and/or secondary disclosure.

Here, additional cells have been suppressed, despite the fact that their counts were not below the threshold. Exactly which additional cells are suppressed is chosen by the researcher (unless specified under dataset- or project-specific SDC rules). This secondary suppression ensures that each row and each column contains at least **two** suppressed values, preventing recalculation of the suppressed values using the totals and other values in the row or column.

Note that the researcher may use the same or different symbols or letters to indicate primary versus secondary suppression, provided that the choice does not enable the reader to crack the suppression.

Additionally, note that secondary suppression only applies to counts and statistics (e.g., percentages, means, ratios) that **enable recalculation** of counts and statistics that have undergone primary suppression. Therefore, a table such as this:

Ethnicity	N	Mean score
Asian	46	8.4
Black	12	6.1
Mixed	21	7.9
Other	6	9.9
White	73	7.2
Total	158	7.4

Could have primary suppression like this (as statistics relating to counts below threshold must be suppressed but the researcher chooses the symbols or letters that indicate suppression):

Ethnicity	N	Mean score
Asian	46	8.4
Black	12	6.1
Mixed	21	7.9

Other	<10	-
White	73	7.2
Total	158	7.4

'-' indicates suppression due to low counts.

However, its secondary suppression could look like this:

Ethnicity	N	Mean score
Asian	46	8.4
Black	<15	6.1
Mixed	21	7.9
Other	<10	-
White	73	7.2
Total	158	7.4

'-' indicates suppression due to low counts.

The researcher has chosen to carry out secondary SDC by suppressing the count associated with the 'Black' ethnicity group. This prevents recalculation of the suppressed count for the 'Other' ethnicity group. To preserve maximum information, the researcher has chosen to suppress this count as '<15' rather than less informative options such as '-', '.', 'SUPP', etc. However, the mean associated with the 'Black' ethnicity group does not need to be suppressed. This is because a) it does not relate to a group whose count is below the threshold and therefore it does not need to have primary suppression applied to it and b) it cannot be used to recalculate the suppressed count or the suppressed mean for the 'Other' ethnicity group and therefore it does not need to have secondary suppression applied to it.

Finally, note that secondary suppression rapidly gets very complicated to both implement and check if the output is extensive and/or complex, e.g., it contains multiple inter-related tables or tables with a hierarchical relationship to each other. In this instance, we strongly suggest **not** using secondary suppression as the secondary SDC method. Instead, we strongly recommend using rounding instead.

6.3.3 Example: rounding

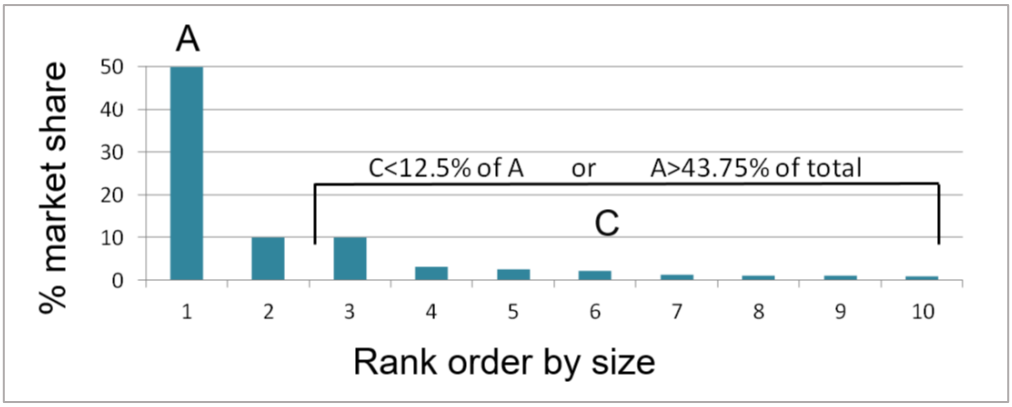
A third method of secondary SDC is rounding. As noted above, this is particularly useful when the secondary disclosure is extensive and/or complex. As an example, we will use the table created in section 6.1.1 again. Rounding is shown in section 6.1.2. Rounding is applied to **all** counts in the table, ensuring consistency and thus preventing output checker and reader confusion.

7. Dominance

Dominance can exist as a potential disclosure risk when there is a particularly large unit of data within a sample. There are two ways of classifying dominance within the SRS (other Trusted Research Environments may use other definitions):

1) If your **largest** entity is **>43.75%** of the total figure and/or
2) If the entities **except the largest two** represent **<12.5%** of the largest entity.

This can be shown graphically, as follows:



Dominance is rare and is often incredibly hard to spot at the output checking stage. However, if it is identified, standard SDC practices are the best way to deal with dominance – i.e., redesign, round, suppress, etc. If any values are suppressed, secondary SDC must be applied to prevent secondary disclosure of the suppressed values by differencing – see section 6.3.

7.1 Example: reformatting

This table of pay data, created in a project that has custom rules permitting outputs at educational institution level, does not have any immediately obvious disclosure issues:

Employment classification	University 1		University 2		Secondary school 1		Secondary school 2	
	N	Mean	N	Mean	N	Mean	N	Mean
Management	45	98,940	10	180,420	15	67,093	x	x
Teacher	158	45,302	45	46,349	92	39,023	15	38,730
Assistant	219	26,823	34	26,392	48	24,567	x	x
Other	80	24,509	18	26,781	26	23,201	x	x

'x' indicates suppression due to low counts.

The dominance in this table is not immediately obvious. It can be identified by looking at the mean pay for 'Management' for University 2. The mean pay for this group is much higher than for 'Management' for University 1, despite having 35 fewer staff in this employment classification in University 1 than University 2. Therefore, even without looking at the record-level data, we know that there will be some very high earners skewing the mean pay for this employment classification for University 2. Topic-specific knowledge can help here: skewing of pay data is likely to be more common and more extensive in smaller universities, as pay scales for upper management are much higher than for departmental chairs. Any university (regardless of size) needs a certain number of upper management positions, but the number of departmental chairs will generally be proportional to the size of the university.

The best way to fix dominance of this nature is to reformat the output. For example:

Employment classification	Sampled universities		Sampled secondary schools		Whole sample	
	N	Mean	N	Mean	N	Mean
Management	55	113,755	23	75,424	78	102,452
Teacher	203	45,534	107	38,982	310	43,273
Assistant	253	26,765	52	25,058	305	26,474
Other	98	24,926	28	23,485	126	24,606

Here, the data has been pooled by type of educational institution, rather than reporting at an individual educational institution level. This alleviates the dominance and therefore reduces the risk of disclosure.

8. Statistics

Various statistics may be requested for output, ranging from routine descriptive statistics (e.g., mean, median, mode, standard deviation, range, percentiles, minimums, maximums) to more specialist statistics (e.g., variance, covariance, kurtosis, skewness, hypothesis testing, concentration ratios, odds ratios). These should all be reported based on the threshold, as follows.

Note that in all cases the relevant count is the number of data subjects used to calculate the statistic, not the number of data subjects in the group overall – sometimes information for some variables is only available for a subset of the data subjects and this should be considered when determining the underlying count for a given statistic.

8.1 'Safe' statistics

Standard deviation, higher moments of distributions (e.g., variance, covariance, kurtosis, skewness) and concentration ratios (including the Herfindahl-Hirschman Index (HHI)) are considered 'safe' statistics. This is because the mathematical complexity of their production ensures that it is very difficult to use them to meaningfully disclose anything about individual data subjects without considerable additional information.

Generally, these 'safe' statistics are not considered to be disclosive provided that the underlying counts are a) stated and b) meet or exceed the threshold.

If the researcher has any other statistic that they wish to output that is not listed above but they think constitutes a 'safe' statistic due to its mathematical complexity, they must provide an explanation of its 'safety'. This information should be provided via annotating the output file and/or adding a note in the description section of the Output Request form.

8.2 'Unsafe' statistics

Other statistics are considered 'unsafe' and therefore may require more accompanying information or may be unable to be cleared in the majority of cases.

8.2.1 Mean

Means represent an increased disclosure risk as they are easily calculated: one simply needs the value held by each individual in the group and the count of individuals in the group.

The same rules apply to means as other statistics: i.e., underlying counts must be a) provided and b) meet or exceed the threshold.

Due to the simplicity of their calculation, means may be associated with secondary disclosure. Outputs containing means should be thoroughly checked to exclude this possibility – see section 6.3 for more details.

8.2.1.1 Example: suppression

Means may be presented in text or in tables or graphs. For an example of a mean presented as a graph, see sections 9.1.1 and 9.1.2. Here is an example of a mean presented in a table:

Grant status	No. employees (mean)	Turnover (mean change)
Grant A	128	1.4
Grant B	73	0.3
Grant A & B	114	2.2
No grant	106	0.2
Unknown	131	0.7

These means, like all means, cannot be cleared without their underlying counts, which are:

Grant status	Counts underlying No. employees (mean)	Counts underlying Turnover (mean change)
Grant A	34	34
Grant B	26	26
Grant A & B	11	9
No grant	11,917	11,917
Unknown	5	5

This reveals that several of the means are calculated from groups of firms whose count is below the threshold (highlighted in yellow to make them clearer). Where counts of a group are below threshold, the statistic calculated from this group must be suppressed:

Grant status	No. employees (mean)	Turnover (mean change)
Grant A	128	1.4
Grant B	73	0.3
Grant A & B	114	.
No grant	106	0.2
Unknown	.	.

'.' indicates suppression due to low counts.

Additionally, checks should be made to ensure that these means cannot be used to difference suppressed values elsewhere in the output and that other information in the output cannot be used to difference these suppressed means – see section 6.3.

8.2.2 Percentages

Similarly to means, percentages represent an increased disclosure risk as they are easily calculated: one simply needs the count of a subgroup (i.e., numerator) and the count of the group that subgroup belongs to (i.e., denominator).

The same rules apply to percentages as other statistics: i.e., underlying counts must be a) provided and b) meet or exceed the threshold. Note that 'underlying counts' refers to the numerator of the percentage, not the denominator.

Due to the simplicity of their calculation, percentages are frequently associated with secondary disclosure. Outputs containing percentages should be thoroughly checked to exclude this possibility – see section 6.3 for more details.

8.2.3 Weighted counts

Similarly to means and percentages, weighted counts represent an increased disclosure risk as they are often easily calculated: e.g., their calculation may be as simple as multiplying the count by a constant.

The same rules apply to weighted counts as other statistics: i.e., underlying counts must be a) provided and b) meet or exceed the threshold. Note that 'underlying counts' refers to unweighted counts. Be aware that weighting in and of itself is **not** sufficiently protective of low counts as it is often possible to reverse the weighting using a combination of reported methodology, weighted counts and/or unweighted totals. These may be present in the output or elsewhere in the public domain.

Weighted counts may also be associated with secondary disclosure. Outputs containing weighted counts should be thoroughly checked to exclude this possibility – see section 6.3 for more details.

8.2.4 Mode, minimums and maximums

Mode, minimums, and maximums represent an increased disclosure risk as they typically relate to individual data subjects. Therefore, the mode, minimums and maximums should not be included in outputs except in a few circumstances, where they may not be disclosive, e.g.:

- When the value is held by at least threshold number of data subjects (e.g., a sample has a modal depression score of 2 on a 5-point Likert scale, where the number of data subjects that had a score of 2 is a) stated and b) meets or exceeds the threshold).
- When the value is structural as the variable's range is limited (e.g., a minimum of zero and a maximum of 100 for a variable that is reported as a percentage).

If the researcher wishes to clear modes, minimums and/or maximums, they must demonstrate why the modes, minimums and/or maximums in the output are not disclosive. This information should be provided via annotating the output file and/or adding a note in the description section of the Output Request form.

8.2.4.1 Example: suppression

Modes, minimums and maximums are generally not able to be cleared except in the few circumstances described in section 8.2.4. For example:

	Minimum	Maximum
Age	11	16
GCSE English score (%)	0	100
GCSE Mathematics score (%)	0	100
GCSE History score (%)	1	97

Here, the minimum and maximum for 'GCSE English score' and 'GCSE Mathematics score' may be released – as the scores are percentage scores, a minimum of zero and maximum of 100 are structural rather than informative, regardless of how many individuals hold them. However, the minimum and maximum for 'Age' and 'GCSE History score' are not structural and therefore cannot be cleared without the count of data subjects holding each value:

	Minimum	Maximum	Count underlying minimum	Count underlying maximum
Age	11	16	1	512
GCSE English score (%)	0	100	N/A	N/A
GCSE Mathematics score (%)	0	100	N/A	N/A
GCSE History score (%)	1	97	4	1

Some of the counts underlying the 'Age' and 'GCSE History score' minimum and maximum are below threshold (highlighted in yellow to make them clearer) they must be suppressed as follows:

	Minimum	Maximum	Count underlying minimum	Count underlying maximum
Age	-	16	-	512
GCSE English score (%)	0	100	N/A	N/A
GCSE Mathematics score (%)	0	100	N/A	N/A
GCSE History score (%)	-	-	-	-

'-' indicates suppression due to low counts.

This table illustrates well how under the circumstances that modes, minimums and maximums are permitted, they are often no longer useful from a research perspective. This is why it is generally advised that minimums and maximums be simply avoided.

It should be noted that there are two different interpretations for this table; the minimum and maximum for English and Maths could be considered structural but for History the minimum and maximum represent the exact scores the students achieved. When presenting both together the implication is that 0 and 100 were the minimum and maximum score achieved by a certain number of students, not of all possible scores the students could achieve, and thus could be considered disclosive so underlying counts and suppression might be needed. To avoid reader confusion, minimum and maximum should refer to the same concept, e.g. the possible scores to get in GCSE or the scores students actually achieved, and this should be clear in the table description.

8.2.4.2 Example: rounding

If the mode, minimum and/or maximum are especially necessary for a project's research goals, it is sometimes possible to use rounding to enable them to meet one of the exceptions under which they may be released. For example:

	Firm distance from innovation centre (miles)
Mode	43.1
Minimum	0.1
Maximum	192.9

These statistics cannot be cleared without the count of data subjects holding each value, which are as follows:

	Firm distance from innovation centre (miles)	Underlying firm-level count of statistic
Mode	43.1	2
Minimum	0.1	4
Maximum	192.9	1

These counts are all below the threshold. Therefore, these statistics cannot be cleared. However, by rounding the output, the counts held by some of these statistics now meet or exceed the threshold:

	Firm distance from innovation centre (to nearest 5 miles)	Underlying firm-level count of statistic
Mode	45	12
Minimum	0	31
Maximum	190	1

However, as shown above, sometimes even rounding is not sufficient to make this type of statistic suitable for clearance – in this example, the count underlying the maximum (highlighted in yellow to make it clearer) is still below the threshold even after rounding. Therefore, this statistic must be suppressed:

	Firm distance from innovation centre (to nearest 5 miles)	Underlying firm-level count of statistic
Mode	45	12
Minimum	0	31
Maximum	SUPP	SUPP

'SUPP' indicates suppression due to low counts.

8.2.5 Medians, quartiles, deciles and percentiles

Medians, quartiles, deciles and percentiles represent an increased disclosure risk as they typically relate to individual data subjects. This is particularly risky when the sample is small and/or when the

percentile is further from the median (as under these circumstances the statistic is more likely to represent an outlier in the dataset). Therefore, the following rules apply:

- Median (aka quartile 2 (Q2) or 50th percentile) should be suppressed if the total count is less than **twice** the threshold. As the median corresponds to half of the group, each half needs to meet the threshold, the total needs to be at least twice the threshold,. Therefore, using a threshold of 10 there would need to be at least 20 subjects in the total group.
- Upper and lower quartile (aka quartile 3 (Q3) and quartile 1 (Q1) or 75th and 25th percentiles) should be suppressed if the count of the group is less than **four times** the threshold (as each quartile corresponds to one quarter of the group). Therefore, using a threshold of 10 would need to be at least 40 people in the total group.
- Deciles/percentiles should be suppressed based on the count of the group that they correspond to, e.g.:
    - Deciles correspond to tenths of the group, so should be suppressed if the count of the group is less than **ten times** the threshold (as each decline corresponds to one tenth of the group).
    - The 1st and 99th percentile correspond to hundredths of the group, so should be suppressed if the count of the group is less than **one hundred times** the threshold (as each percentile corresponds to one hundredth of the group).
    - Etc.

Note that these rules apply regardless of the format that the medians, quartiles, deciles and/or percentiles are presented in. For examples of these statistics graphed, see section 9.4.

8.2.5.1 Example: suppression

Medians, quartiles, deciles and percentiles are often presented in tables, as below, though these rules apply equally no matter the format they are presented in.

	Median	LQ	UQ	N
All	47	34	58	100
Male	47	20	68	46
FSM	52	30	63	65
SEN	46	19	62	36
SEN (EHCP)	22	14	49	11

The best way to handle the disclosure issues in this example is via suppression – i.e., suppressing the median where the count is less than twice the threshold and suppressing quartiles where the count is less than four times the threshold. For example:

	Median	LQ	UQ	N
All	47	34	58	100
Male	47	20	68	46
FSM	52	30	63	65
SEN	46	.	.	36
SEN (EHCP)	.	.	.	11

'.' indicates suppression due to low counts.

8.2.5.2 Example: reformatting

In some circumstances, reformatting is a suitable way to handle a disclosure risk caused by percentiles. For example:

	Attendance (%)			
	Median	1st percentile	99th percentile	N
Before scheme	86	32	100	971
After scheme	93	64	100	964

In this table, the 1st and 99th percentiles are present despite the count of each group being less than one hundred times the threshold (highlighted in yellow to make them clearer). Therefore, it cannot be cleared. However, as the count of the group is still quite high (and greater than twenty times the threshold), reformatting the table to use vigintiles instead of percentiles would make it suitable for clearance:

	Attendance (%)			
	Median	5th centile	95th centile	N
Before scheme	86	40	97	971
After scheme	93	67	98	964

8.2.6 Ratios, including odds ratios

Similarly to means, percentages and weighted counts, ratios (including odds ratios) represent an increased disclosure risk as they are typically easily calculated.

The same rules apply to ratios as other statistics: i.e., underlying counts must be a) provided and b) meet or exceed the threshold. Note that for ratios there are several underlying counts that are relevant and all numerators for these should be provided.

Due to the simplicity of their calculation, ratios may be associated with secondary disclosure. Outputs containing ratios should be thoroughly checked to exclude this possibility – see section 6.3 for more details.

(Note: concentration ratios, such as the Herfindahl-Hirschman index, have much more complex methodology than the ratios described in this section. Therefore, concentration ratios are considered 'safe' statistics and are covered in section 8.1.)

8.2.6.1 Example: suppression

It can sometimes be a little complicated to determine which are the appropriate underlying counts for a ratio. For example, in this table reporting the odds ratio of cancer for smokers versus non-smokers:

	Relative risk	Odds ratio
Lung cancer	9.96	10.80
Liver cancer	3.73	3.85
Bladder cancer	4.43	4.90

Relative risk (also called risk ratio) and odds ratio are both calculated from the number of cancer cases within the 'smoker' group and the number of cancer cases within the 'non-smoker' group. Therefore, both of these numerators are relevant as underlying counts for these ratios:

	Smoker			Non-smoker		
	Cancer	No cancer	Total	Cancer	No cancer	Total
Lung	786	8,424	9,210	100	11,574	11,674
Liver	387	8,823	9,210	132	11,600	11,732
Bladder	39	288	327	6	217	223

The relative risk and the odds ratio for bladder cancer are derived from a count below threshold (highlighted in yellow to make it clearer). Therefore, these two statistics must be suppressed:

	Relative risk	Odds ratio
Lung cancer	9.96	10.80
Liver cancer	3.73	3.85
Bladder cancer	-	-

'-' indicates suppression due to low counts.

9. Graphs

Graphic representations of data can present a high risk of disclosure, particularly if the methods of data presentation are used to show the distributions of a value (e.g., histograms) or if bars, points or lines relate to a single observation or a single data subject (e.g., scatter plots).

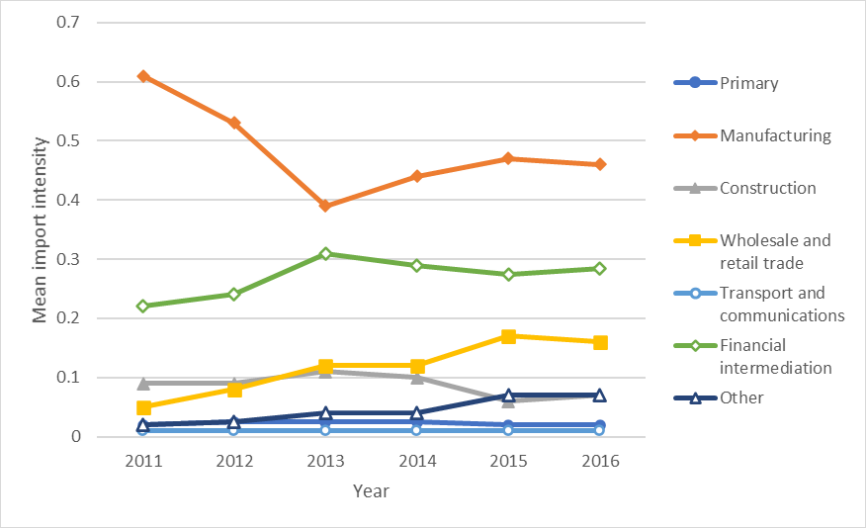
Graphic representations of data are subject to the same SDC methodology as tables. Therefore, all graphic representations of data should be presented with their underlying unweighted counts. These underlying counts must meet or exceed the threshold (except in cases where zeros are structural – see section 6.2). A convenient way to do this is to provide the underlying counts in table(s) in supplementary Excel file(s). If you do this, you should clearly indicate (e.g., in the Output Request form) where to find the underlying counts for each graph, citing file names and sheets, pages, lines and/or cells within files, as applicable.

9.1 Line graphs

Line graphs indicate the strength and direction of a relationship between two or more variables – one of which is often time. Generally, line graphs are not considered to be disclosive provided that the underlying counts are a) stated and b) meet or exceed the threshold (except in cases where zeros are structural – see section 6.2).

9.1.1 Example: suppression

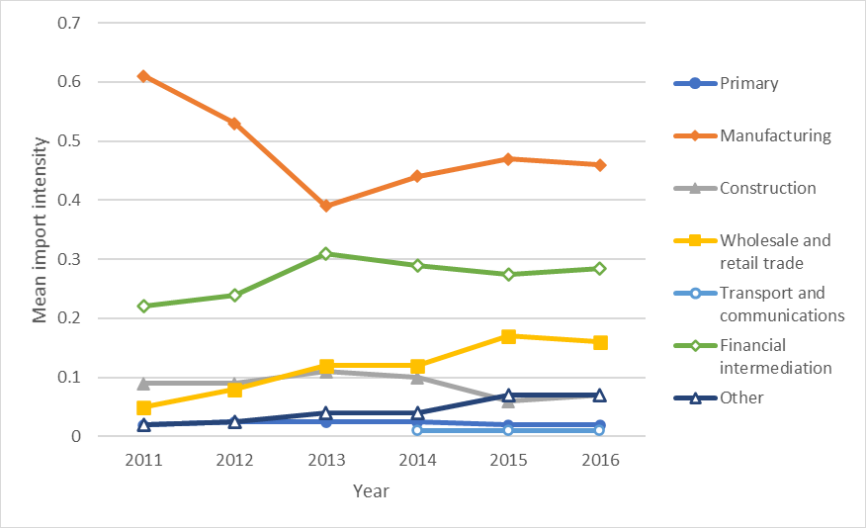
Line graphs may not seem disclosive at first but may contain a variety of disclosure risks. For example:



This line graph, like all graphs, cannot be cleared without its underlying counts, which are:

Industry	2011	2012	2013	2014	2015	2016
Primary	69	89	88	76	170	157
Manufacturing	2764	2149	1570	1756	3863	3850
Construction	395	377	480	418	382	410
Wholesale and retail trade	209	319	487	494	1314	1301
Transport and communications	9	8	6	35	84	111
Financial intermediation	987	973	1223	1182	2198	2364
Other	83	97	137	142	409	398

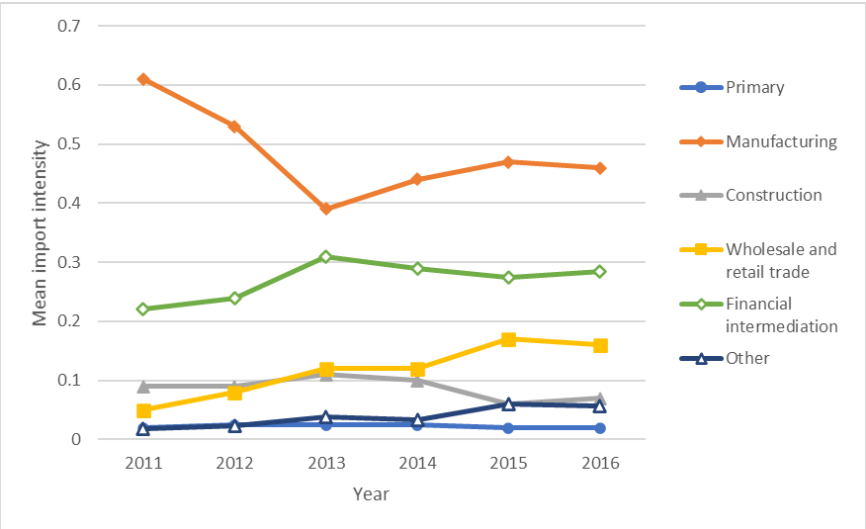
As the underlying counts for 'Transport and communication' for years 2011-2013 are below the threshold, the line graph cannot be cleared. Similarly to the table that underlies it, there are a number of ways to apply SDC to this line graph – for example, the low counts could be suppressed, or the graph could be reformatted by combining business categories and/or by combining reporting years. For more details, see section 6.1. Note that any SDC must be done in the graph as well as the frequency table that shows the graph's counts:



Note: low underlying counts prevent reporting of 'Transport and communications' for 2011-2013.

9.1.2 Example: reformatting

Alternatively, disclosive line graphs may be reformatted: groups may be pooled or dropped entirely. For example, here the 'Transport and communications' group has been merged with the 'Other' group so that all groups now have counts that meet or exceed the threshold:



9.2 Scatter graphs

Scatter graphs also indicate the strength and direction of a relationship between two or more variables. However, each point on a scatter graph typically represents one data subject. It is, therefore, relatively easy to identify an individual data subject and to attribute data to them when using scatter graphs. For this reason, scatter graphs are generally considered to be problematic and are not routinely permitted in outputs.

There are some instances when scatter graphs are not necessarily disclosive. The main example is binned scatter graphs[3], where the graph is formatted so that each point represents a bin containing multiple data subjects. Binned scatter graphs are generally considered non-disclosive provided that the number of data subjects in **each** bin on the graph is a) clearly stated and b) meets or exceeds the threshold.

Some other instances where scatter graphs are not necessarily disclosive include:

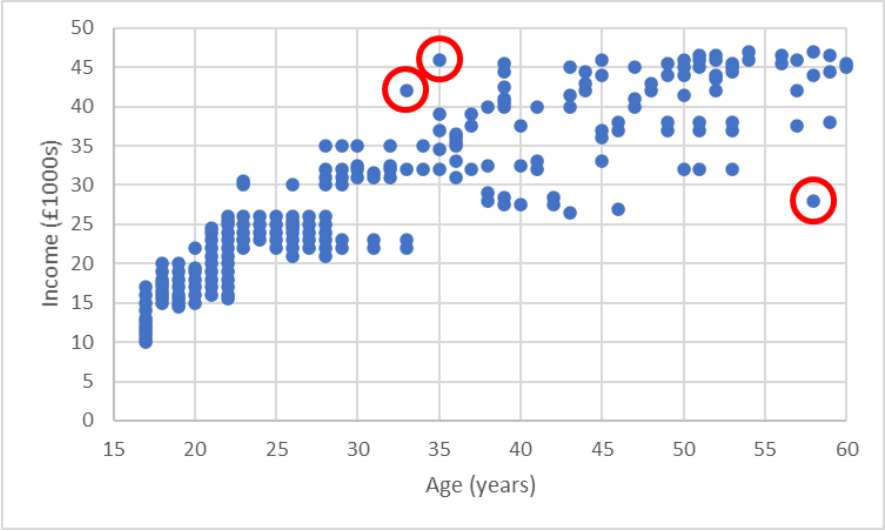
- Scatter graphs of 'safe' statistics. E.g., a scatter plot of the distribution of kurtosis among subsamples, where each subsample met the SDC 'rule of thumb' for reporting kurtosis.
- Scatter graphs of variables, particularly highly derived variables, that could not reasonably be related back to underlying counts except by the researcher. E.g., a scatter plot of predicted turnover by industry for the next 5 years, created using a model designed by the researcher.
- Scatter graphs that use standardised or otherwise transformed statistics, e.g.:
    - Where one or both axes are standardised, e.g., so that the mean is zero and the standardisation is one or so that the minimum is zero and the maximum is 100.
    - Where one or both axes are presented as 'distance from…', e.g., 'distance below median' or 'distance above mean'.
    - Where the scale is suppressed on one or both axes.

  However, note that standardisation and transformation do not, in and of themselves, represent sufficient risk mitigation as mean, median, standard deviation, range, etc. may be presented elsewhere in the output or in the public domain.

However, these and other scatter graphs may only be cleared as an exception.

9.2.1 Example: reformatting

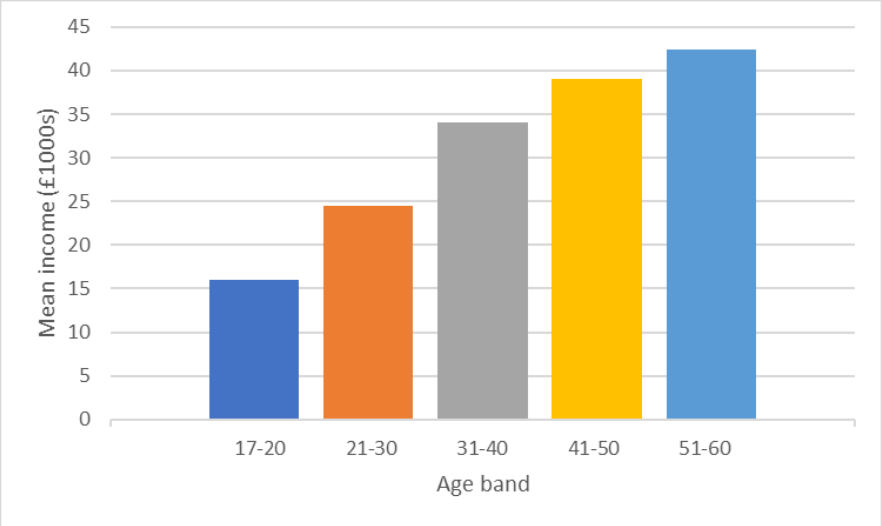
Scatter graphs, as described above, are generally disclosive as each point typically represents a single data subject. For example:



[3] For R users, ggplot2 3.3.0 and later has a bin scale option which may be of use when producing these outputs.

Even though there are no extreme outliers in this graph, each point can be used to provide information about a single data subject. Therefore, it is relatively easy to pick out an individual data subject and attribute data to them. For example, the red circled points (from left to right) show us that the data includes an individual aged 33 earning £42,000 per annum, an individual aged 35 earning £46,000 per annum and an individual aged 58 earning £28,000 per annum.

This demonstrates why the majority of scatter plots are not suitable for clearance. Instead, the data could be reformatted – e.g., as a bar chart, histogram, line graph, frequency table and/or as statistic(s) (e.g., mean). For example:



9.3 Bar charts and histograms

Bar charts represent the relative frequency of observations or variables in relation to one another, whereas histograms display the frequency distribution of a variable with the width of bars representing either class intervals or a single value and the height of bars representing frequency or density. Both of these types of graph, therefore, provide insight into the distribution of the data and can be useful in displaying the shape of the distribution of specific variables.

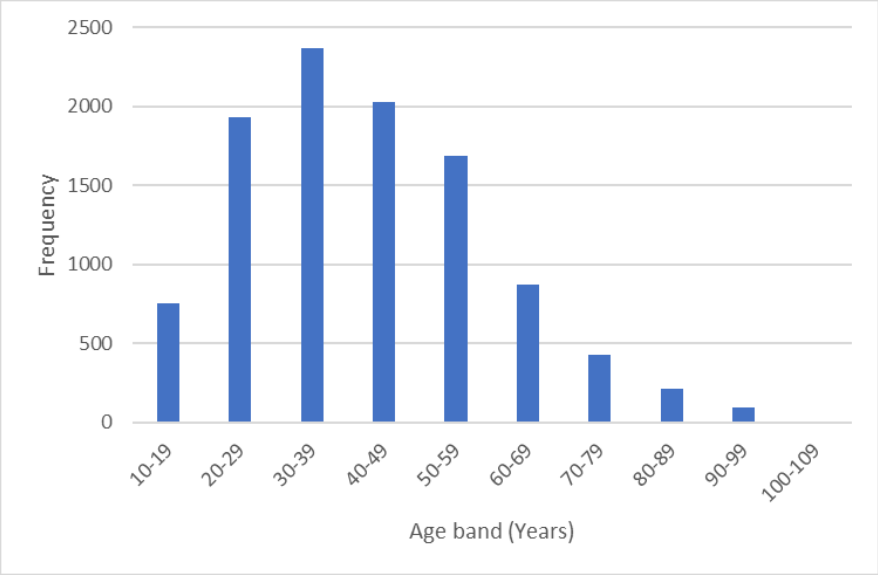
The same mitigating actions that apply to other methods of graphic representation also apply to bar charts and histograms. Generally, bar charts and histograms[4] are not considered to be disclosive provided that the underlying counts are a) stated and b) meet or exceed the threshold (except in cases where low counts are structural – see section 6.2).

Bars in either bar charts or histograms that relate to counts below threshold should be either suppressed or reformatted (e.g., combining categories or rounding) to prevent disclosure.

9.3.1 Example: reformatting

Histograms may also initially seem safe but can contain disclosure risks. For example:

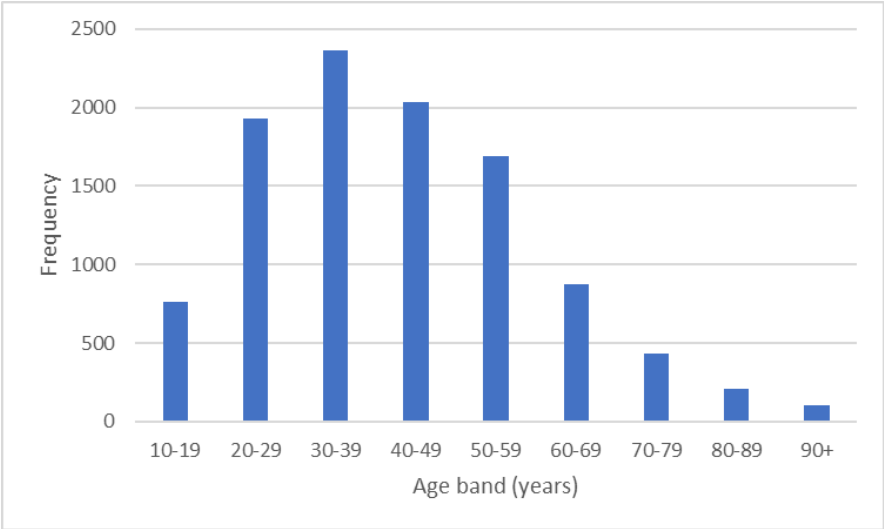
[4] For R users, see footnote 3 regarding the ggplot2 3.3.0 and later bin scale option which may be of use when creating histograms.



This histogram, like all graphs, cannot be cleared without its underlying counts, which are:

Age band (years)	Frequency
10-19	758
20-29	1,932
30-39	2,367
40-49	2,031
50-59	1,689
60-69	876
70-79	432
80-89	211
90-99	97
100-109	4

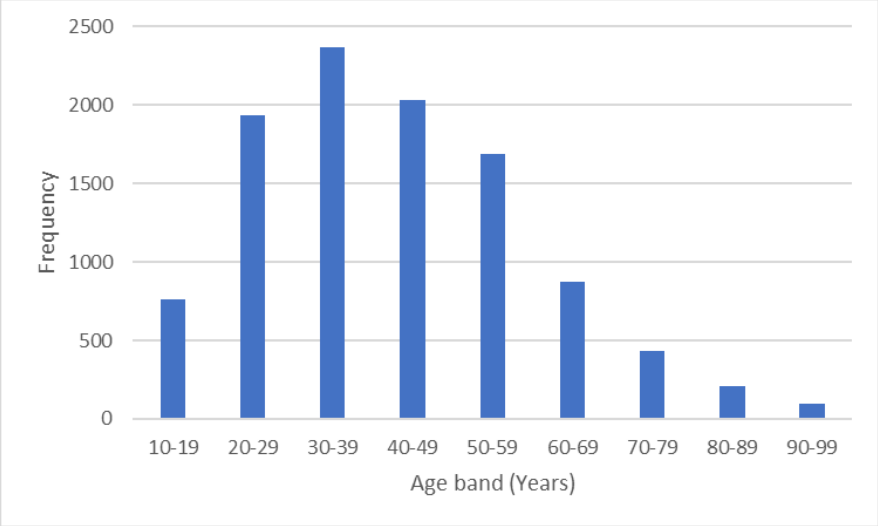
As the underlying counts for the age band '100-109' are below the threshold, the histogram cannot be cleared. Typically, the most suitable way to handle a disclosure issue like this is by reformatting the histogram, combining some of the bars as follows:



As the '90-99' and 100-109' age bands have been combined, there are no longer any bars that represent a group with a count below threshold. Therefore, this histogram meets the SDC 'rules of thumb'.

9.3.2 Example: suppression

Alternatively, in some cases histograms and bar charts that are disclosive can be altered by suppression. For example:



Note: age bands above 99 years have been suppressed due to low underlying counts.

Suppression of bars should only be chosen when it does not risk confusing the reader, e.g., by implying that the category holds a value of zero. It is strongly advised that a note accompany the graph so that it is clear where data has been suppressed

9.4 Boxplots

Boxplots typically have a box showing the upper and lower quartiles and a bar inside the box showing the median. They also typically have whiskers, which may show the minimum and maximum, a multiple of the interquartile range above and below the box, the 2nd and 98th percentiles or other values. If the whiskers are not the minimum and maximum, any outliers beyond the whiskers are typically shown as points. Sometimes the mean is shown as a point. Sometimes the boxplot is notched, to indicate a confidence interval around the mean. Sometimes, dot plots (effectively single-axis scatter plots) are overlaid over boxplots.

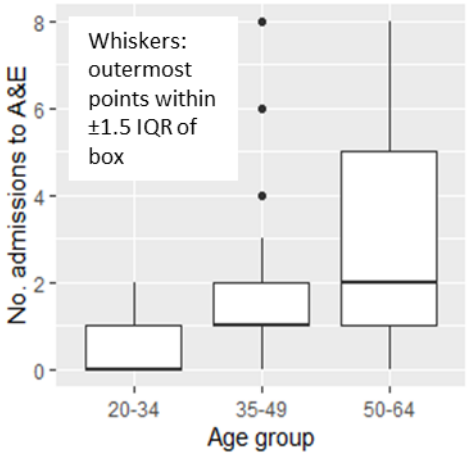
In general:

- Boxplots should be suppressed if the count of the group is less than **four times** the threshold (as boxplots show quartiles).
- The plot should clearly indicate what the whiskers (and any points) show.
- Whiskers should comply with the above guidance relating to minimums, maximums, deciles and percentiles (as applicable). This means whiskers may need to be changed to a different whisker type or suppressed entirely. If they are suppressed, this should be indicated to avoid reader confusion.
- Outliers should be suppressed, as these represent values relating to single data subjects.

- The mean may be shown as a point.
- The boxplot may be notched to show confidence intervals.
- Any dot plots overlaying boxplots should be treated in the same way as scatter plots – i.e., they are generally not permitted. This applies even if the points are jittered (i.e., some random noise has been introduced into the dot plot).

9.4.1 Example: suppression of plots

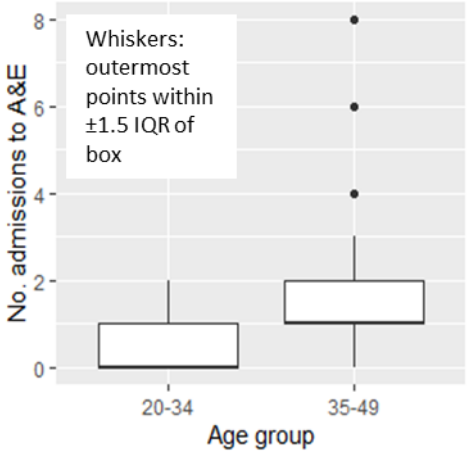
Boxplots can look innocuous at first but may contain a variety of disclosure risks. For example:



This set of boxplots, like all graphs, cannot be cleared without its underlying counts, which are:

Age group	No. admissions to A&E
20-34	104
35-49	124
50-64	39

As boxplots report quartiles and the count for the age group '50-64' is less than four times the threshold, this age group cannot be reported and must be suppressed:

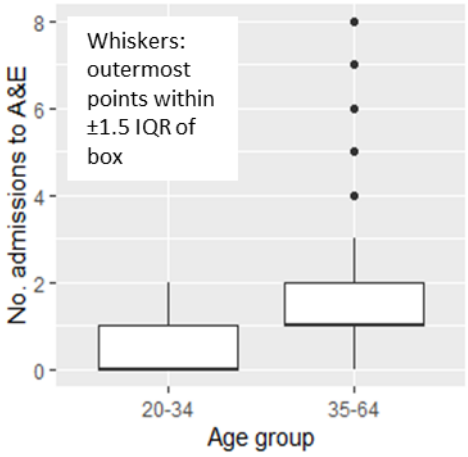


In this instance, this suppression is not sufficient on its own to prevent disclosure. The other issues with this set of boxplots are discussed in sections 9.4.3 and 9.4.4.

Note that if the suppressed boxplot(s) were necessary, reformatting might be a better option – see section 9.3.4.

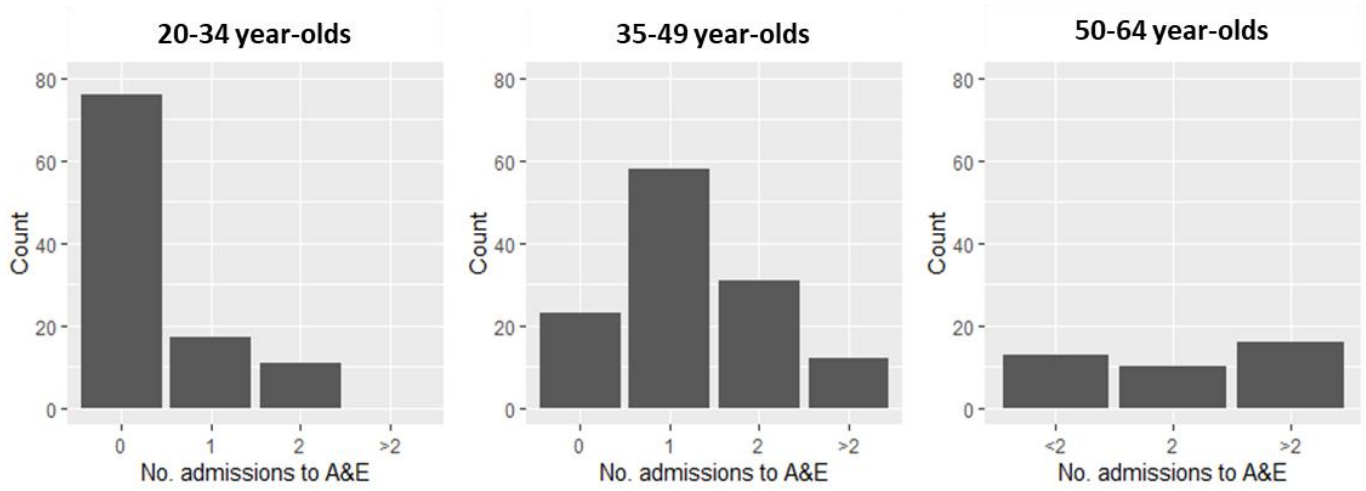
9.4.2 Example: reformatting plots

If boxplots are not permissible, as the group(s) being plotted have a count less than four times the threshold, reformatting can be used as a solution. There are two general methods: pooling groups or converting the boxplot into a histogram with pooled categories. For example, using the set of boxplots from section 9.4.1, the groups could be pooled as follows:



Here, the '35-49' and '50-64' age groups have been merged, creating a group with a count greater than four times the threshold (see table in section 9.4.1). Note that other issues remain with this set of boxplots that are discussed in sections 9.4.3 and 9.4.4.

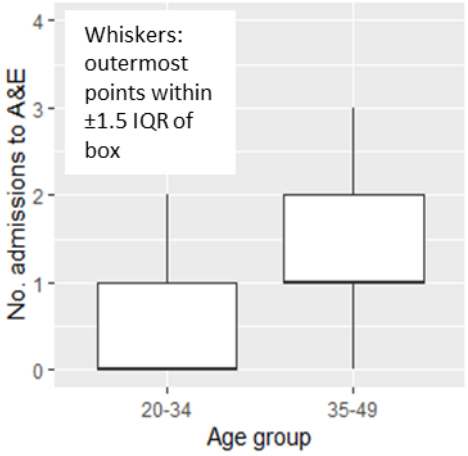
Alternatively, the boxplots could have been reformatted into histograms with pooled categories, e.g.:



Here, each pooled category has a count that meets or exceeds the threshold. Therefore, these histograms meet the SDC 'rules of thumb'.

9.4.3 Example: suppression of outliers

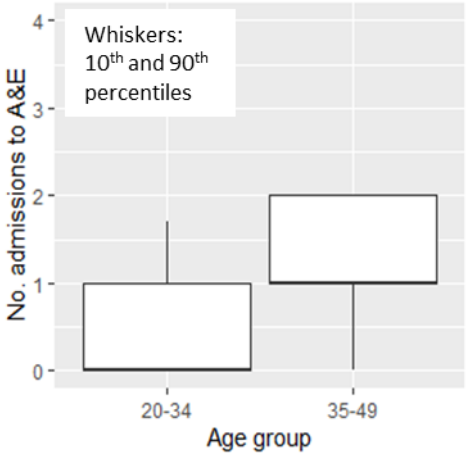
Generally, outliers are not permitted as they represent single data subjects. Therefore, they should be suppressed. For example, using the set of boxplots from section 9.4.1:



Note that the y-axis scale has been changed to avoid implying the extent of the suppressed outliers. However, in this instance, this is still not sufficient on its own to prevent disclosure. The other issue with this set of boxplots is discussed in section 9.4.4.

9.4.4 Example: reformatting whiskers

Boxplot whiskers may be defined in a variety of ways. However, the choice of whiskers on boxplots should be made carefully, to avoid disclosure. For example, the set of boxplots from section 9.4.3 use 'outermost points within ±1.5 interquartile ranges of the box' to define the whiskers – i.e., the whisker definition originally suggested by Tukey when making rules for boxplots. However, the number of data subjects meeting this specification is frequently below threshold or even one. Therefore, whiskers with this definition frequently present a disclosure risk. In this instance, the boxplot should be reformatted to use a different whisker definition, e.g.:



Here, the whiskers have been reformatted so that they are defined as the 10th and 90th percentiles. As both of the age groups have a count at least 10 times the threshold (see table in section 9.4.1), it is fine to use these percentiles. However, note that if particularly small sample sizes are used, the

whiskers may have to be suppressed entirely, in which case, if a boxplot must be used, this should be indicated via an annotation to avoid reader confusion.

9.5 Violin plots

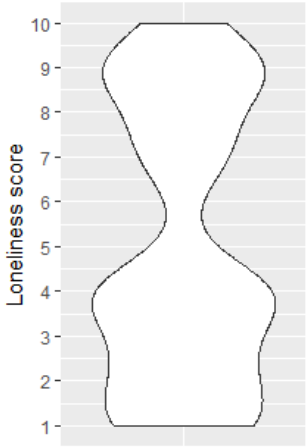
Violin plots show the probability density of all the data, with smoothing applied. They may also have a marker showing the median. Sometimes a small boxplot will overlay the violin plot to show the quartiles. Sometimes the mean is shown as a point.

As violin plots show the full distribution of the data, there is a much higher risk of disclosing information relating to small counts or single data subjects than with a boxplot. The risk is higher if the plotted data are ordinal numbers (e.g., scores on a scale) or continuous integers (e.g., age in years) rather than continuous decimal numbers (e.g., income in decimal pounds). The risk is also higher when the sample size is small compared to the y-axis range and scale (e.g., for a sample of 20 individuals, presenting 'job satisfaction' as a 5-point Likert scale (y-axis of 0-4 in integers) is less risky than presenting 'attendance over last calendar year' as a decimal percentage (y-axis of 0.0 to 100.0)).

In general, violin plots may **only** be cleared as an exception, as described in section 2.

9.5.1 Example: reformatting plots

Violin plots can also look innocuous at first but may contain a variety of disclosure risks. For example:



This violin plot, like all graphs, cannot be cleared (even by exception) without its underlying counts, which are:

Loneliness score	N
1	120
2	90
3	80
4	160
5	10
6	9

7	71
8	60
9	140
10	40
Total	780

As violin plots report the **whole** distribution, the numerators are important (in contrast to boxplots where the denominator is important). The underlying counts demonstrate that the 'pinch point' visible in the graph for a Loneliness score of 6 relates to a group with a count below threshold (highlighted in yellow to make it clearer). Therefore, the violin plot cannot be cleared. Instead, the data could be reported as a boxplot, as described in section 9.4.

Note that the only way to determine if a violin plot is disclosive is by looking at the underlying counts. A 'pinch point' does not conclusively indicate disclosure, nor does the absence of 'pinch points' conclusively indicate a lack of disclosure. For example:

This violin plot does not have any 'pinch points'. However, the underlying counts are:

Loneliness score	N
1	6
2	9
3	8
4	6
5	4
6	9
7	7
8	6
9	3
10	8
Total	66

As with the previous violin plot, this violin plot cannot be cleared. This is because every Loneliness score relates to a group with a count below threshold. Instead, the data could be reported as a boxplot, as described in section 9.4.

10. Regressions and modelling

10.1 Coefficients, margin plots and test statistics

Regressions/models, including their margin plots and test statistics, must be based on degrees of freedom (the number of observations minus the number of parameters being estimated) that are a) stated and b) meet or exceed the threshold. Additionally, regressions/models must not be based on a single unit (e.g., a time series of one business) and must not be saturated. A saturated regression/model is one that reports all interactions of the variables; this can also be called a fully interacted regression/model.

Care should be taken when submitting the unedited outputs of regressions/models carried out by software such as STATA and SPSS, as often other statistics are automatically reported as part of the code output. Some of these, such as kurtosis and skewness, are generally unproblematic – see section 8.1. However, sometimes minimums, maximums and percentiles may be automatically reported by the software program – these are considered on their own merits, following the rules in section 8.2.4.

10.1.1 Example: saturated regression

Saturated regressions are problematic as their coefficients can be summed to give the means of the input variable. For example[5], here a regression has been carried out on 692 records using three binary variables: 'grant_giver', 'survivor' and 'grant_x_surv'. The variable being predicted is 'log_income':

	Coef.	Std. Error	t	p>|t|	95% Conf. Interval	
grant_giver	.6489983	.7343644	0.88	0.377	-.7928660	2.090863
survivor	2.790455	.2385190	11.70	0.000	2.322142	3.258767
grant_x_surv	1.412821	.7896063	1.79	0.074	-.1375061	2.963148
_cons	10.85469	.1839748	59.00	0.000	10.49347	11.21591

All possible combinations of the data have been reported in this regression, i.e.:

- Gave a grant and did not survive: coefficient for 'grant_giver'.
- Did not give a grant and survived: coefficient for 'survivor'.
- Gave a grant **and** survived: coefficient for 'grant_x_surv'.
- Did not give a grant and did not survive: coefficient for the constant ('_cons').

Therefore, this regression is saturated – i.e., it is overfitted because every possible combination of states is represented by the coefficients. When a regression or model is saturated, the coefficients exactly correspond with the means:

Mean log_income		grant_giver	
		No	Yes
survivor	**Dead in 2015**	10.85469	11.50369
	Alive in 2015	13.64514	15.70696

[5] This example is kindly provided by Felix Ritchie, University of the West of England (UWE).

I.e.:

- Did not give a grant:
  - Did not survive: mean = coefficient of the constant.
  - Survived: mean = coefficient of the constant + coefficient of 'grant_giver'.
- Gave a grant:
  - Did not survive: mean = coefficient of the constant + coefficient of 'survivor'.
  - Survived: mean = coefficient of the constant + coefficient of 'grant_giver' + coefficient of 'survivor' + coefficient of 'grant_x_surv'.

As the output of a saturated regression is, effectively, means, if the researcher wishes to have the regression output cleared, they must provide information suitable for clearing means – see section 8.2.1. However, as saturated regressions are typically overfitted and therefore unlikely to have analytical value, it is best to use a different analytical method instead in this circumstance.

10.2 Residuals

Residuals may be plotted or statistics may be calculated from them, e.g., standard deviation. Statistics of residuals are treated using the same rules as if the statistics were of a group of data subjects – see section 8. Plots of residuals are treated the same as any other scatter graphs – i.e., not permitted except in a few circumstances – see section 9.2. If residuals must be plotted, it is preferable to present only the line of best fit on the graph or you can provide a description of the plot in words instead.

11. Maps and spatial analysis

11.1 Maps

Maps are a useful way of showing data that has a geographic component. However, maps can represent a significant disclosure risk. Therefore, choropleth (aka 'heat') maps are strongly encouraged as an alternative to data point maps to avoid disclosing information about single data subjects.

Generally, choropleth maps are not considered to be disclosive provided that the underlying counts are a) stated and b) meet or exceed the threshold (except in cases where low counts are structural – see section 6.2). Choropleth maps that do not meet this requirement should be reformatted to prevent disclosure, e.g., by using a different geography level or choosing another method of visual representation.

Data point maps may only be cleared as an exception, as described in section 2. If an exception of this sort is requested, the researcher should provide clear details outlining what each point represents. Care should be taken that instances of differencing or dominance cannot be inferred through the map itself, by comparing the map to other files in the same output or by comparing the map to other publicly available sources. The researcher should also refer to the guidance in section 11.2.

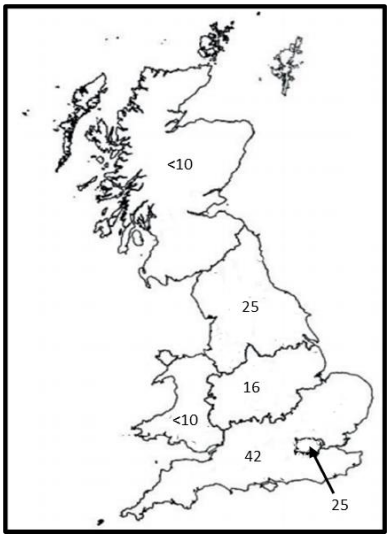
11.1.1 Example: reformatting

Data point maps are generally not permitted as each point typically relates to a single data subject. For example:

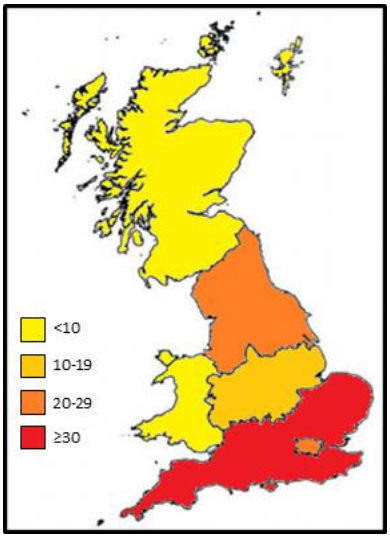
Figure 1: Map showing new cases of disease X recorded in Great Britain in the last 12 months.

Each point relates to a single person with disease X. Although it is difficult to pinpoint exact locations of each case, it is still possible for the individuals to identify themselves and for other individuals to narrow down identities through social networks, news reports and other publicly available information. Therefore, this map represents a significant disclosure risk.

The map could be reformatted so that the data is reported as number of cases per geographic area rather than as single points. For example:



Alternatively, the map could be reformatted as a heat map. For example:



11.2 Geographies

In instances where data is reported based on geographies, care should be taken that the cells in the table, bars/lines/points on a graph and/or points/areas on a map do not relate to low-population density areas such as industrial sites, hamlets, small coastal towns or islands as these could still be considered disclosive. This is particularly true if using Census Output Areas (OAs), which have a minimum count of 40 resident households and 100 resident people in England, Wales and Northern Ireland and a minimum count of 20 resident households and 50 resident people in Scotland.

The geographic level for reporting should be carefully chosen, based on SDC concerns as well as research need.

12. Code files

Code should be reviewed carefully for embedded disclosive data, i.e., low counts, school names, etc. These are most commonly found in hard-coded data, researcher comments and code that discloses information about a single record (e.g., by filtering it using overly specific terms). This approach should be taken with all coding languages.

With code, there is the option of using the Code category of clearance. Therefore:

- If code contains disclosive data, it will be rejected.
- If code contains non-disclosive data, it may be cleared with Publication or Pre-Publication clearance, depending on the researcher's preference.
- If code does not contain **any** SRS data, it may be cleared with Code clearance. This clearance has a much more concise disclaimer and therefore may be preferable, e.g., if the researcher wishes to re-use the code in another SRS project or wishes to make their code public for transparency purposes.

The SDC 'rules of thumb' outlined elsewhere in this document apply to data, regardless of whether it is in code or in a more typical format. Therefore, use these rules when applying SDC to data in code.

12.1 Example: Hard-coded data in code

This code contains hard-coded data:

```
1  # Example 1a - one way to hard-code a table of data (by column)
2
3  age <- c(2, 14, 22, 5, 19, 12, 7, 11)
4  sex <- c("F", "F", "M", "F", "M", "M", "F", "M")
5  cysticfibrosis <- c("Y", "N", "N", "Y", "N", "N", "N", "Y")
6
7  CFstudy <- data.frame(age, sex, cysticfibrosis)
8
```

The code contains the following instructions:

- Rows 3-5 specify the columns for the table.
- Row 7 pulls the columns into a table.

The hard-coded data in rows 3-5 is disclosive – it contains record-level data, as each number/letter in the lists in these rows relate to a single data subject.

This code also contains hard-coded data:

```
14 # Example 1b - another way to hard-code a table of data (by row)
15
16 respondent01 <- c(2, 0, 0)
17 respondent02 <- c(14, 0, 1)
18 respondent03 <- c(22, 1, 1)
19 respondent04 <- c(5, 0, 0)
20 respondent05 <- c(19, 1, 1)
21 respondent06 <- c(12, 1, 1)
22 respondent07 <- c(7, 0, 1)
23 respondent08 <- c(11, 1, 0)
24
25 CFstudy2 <- rbind(respondent01, respondent02, respondent03, respondent04,
26                   respondent05, respondent06, respondent07, respondent08)
27 colnames(CFstudy2) <- c("age", "sex", "cysticfibrosis")
28
```

The code contains the following instructions:

- Rows 16-23 specify the rows for the table.
- Rows 25-26 pull the rows into a table.
- Row 27 re-names the columns of the table.

The hard-coded data in rows 16-23 is disclosive – like the example above, it contains record-level data, as each number in the lists in these rows relate to a single data subject.

12.2 Example: Disclosive comments in code

This code contains comments which contain data:

```
34 # Example 2a - disclosive comments
35
36 hesa_sample <- hesa_sample %>%
37   dplyr::arrange(age) %>%
38   dplyr::filter(!gender == "other")
39 # Filtered out 4 rows
40
```

The code contains the following instructions:

- Row 36 specifies the data that the researcher is working on, which is called ‘hesa_sample’.

- Row 37 sorts the data by age.
- Row 38 filters the data to select rows where the 'gender' variable is not 'other' (therefore removing rows where 'gender' **is** 'other').
- Row 39 contains a researcher comment, stating "Filtered out 4 rows".

The comment in row 39 tells us that there are 4 people in the HESA sample who have 'other' gender. As this count is below threshold, this comment is disclosive.

12.3 Example: Data table in code

This code also contains comments which contain data, though in a different format to the previous example:

```
46  # Example 2b - table in comments
47
48  LFS_sample = LFS_sample %>%
49    dplyr::count(employed_year, ethnicity) %>%
50    dplyr::group_by(employed_year)
51
52  #  ____________________________________
53  # |employed_year | ethnicity | count|
54  # |2020          | White     | 2385 |
55  # |2020          | Black     | 1973 |
56  # |2020          | Mixed     | 672  |
57  # |2020          | Asian     | 1591 |
58  # |2020          | Other     | 342  |
59  # |2021          | White     | 1862 |
60  # |2021          | Black     | 1002 |
61  # |2021          | Mixed     | 451  |
62  # |2021          | Asian     | 1298 |
63  # |2021          | Other     | 279  |
64  #  ____________________________________
65
```

The code contains the following instructions:

- Row 48 specifies the data that the researcher is working on, which is called 'LFS_sample'.
- Row 49 counts the number of data subjects in 'LFS_sample' by the variables 'employed_year' and 'ethnicity'.
- Row 50 groups the counts into a frequency table by 'employed_year'.
- Rows 52-64 contain a researcher comment, which appears to be the frequency table created by lines 48-50 copy-pasted into the code as a comment.

This particular example is not disclosive, as all of the counts in the frequency table exceed the threshold. However, other similar instances of frequency tables in code may be disclosive, depending on their counts and the threshold being used. Additionally, this output could only be given Pre-Publication or Publication clearance, not Code clearance, as it contains SRS data. It is not best practice for code files to include data. Therefore, we recommend that you save any data in a separatee log file instead of introducing data into your code as comments.

12.4 Example: Overly specific code

This code is overly specific, resulting in disclosure of information about a single data subject:

```
79  # Example 4 - overly specific code that discloses information about a single record
80
81  SISclean <- SIS %>%
82    dplyr::filter(!comment == "Please remove my son James from this survey!!! Mrs. Sarah Chambers")
83
```

The code contains the following instructions:

- Row 81 specifies the data that the researcher is working on, which is called 'SIS'.
- Row 82 filters the data to select rows where the 'comment' variable is not 'Please remove my son James from this survey!!! Mrs. Sarah Chambers' (therefore, removing rows where 'comment' **is** 'Please remove my son James from this survey!!! Mrs. Sarah Chambers').

The code in row 82 is overly specific – we know that this comment is present in the dataset and therefore we gain information about a single data subject, including their name (James, probably James Chambers), their relative's name (Mrs. Sarah Chambers) and how that relative is related to the data subject (James is her son). This is disclosive.