SLS-DSU
Scottish Longitudinal Study
Development & Support Unit

SLS Outputs Guidance Document

Document Control:

Name:	SLS Outputs Guidance Document
Version:	3.5
Date Issued:	18/04/2023
Authors:	Including: Susan Carsley, Lee Williamson, Helen Corby, Tom Clemens, Dawn Everington, Lynne Forrest, Gillian Raab, Angela Fallon
Comments to:	sls@ed.ac.uk

Version Control:

Version:	**Date:**	**Comment/ Update**	**Authors**
3.1	10/03/21	Intermediate output form in annex B updated	LF
3.2	12/03/21	Removed references to Intermediate Output clearance spreadsheet	LF
3.3	01/07/21	Small update for adding to SLS website, and in line with ONS LS changed references to intermediate and final to pre-publication and publication output.	LW
3.4	13/07/21	Guidance on synthetic data added as Annex D	LF
3.5	24/02/23	Updated adding in info/text from PHS online documentation	LW & HC

SLS Outputs Guidance Document[1]

On behalf of the Data Controllers, the SLS-DSU and NRS are responsible for managing the risk of directly or indirectly re-identifying any individuals. One of the final steps in this process is to undertake Statistical Disclosure Control (SDC) on all outputs requested for release from the SLS Safe Setting. As such, at the end of your visit, you may not take away any uncleared output.

The SLS-DSU operates a 2-stage approach to cleared outputs:
Stage (1) Pre-publication Output (ie non-disclosive outputs) and ***Stage (2) Publication Output*** (ie anything to be shared beyond the immediate team) - please familiarise yourself with both types as detailed in this guidance document. Further, please do reread the *Annex to the SLS Undertaking Form* which is reproduced at the end of this document which cites immediate sanctions to be applied in the case of breach of the conditions of the undertaking and Annex A the SLS Disclosure Control Protocol.

General Guidance – Summary

Stage (1) Pre-publication Outputs may only be requested in order to produce Stage (2) Publication Outputs. The SLS-DSU no longer offers intermediate outputs (instead, the creation of synthetic data is offered to projects).

If you require Stage (1) Pre-publication Outputs to produce Stage (2) Publication Outputs please **apply** SLS statistical disclosure control (SDC) rules and clearly describe/label your output e.g. give details of variables you have derived and the samples used. Complete the Stage (1) Pre-publication Outputs form (given in Annex B) and email your SLS support staff to describe what it is.

In summary, the main SDC rules for clearing (1) Pre-publication Outputs are:

- These are not intermediate outputs, these are pre-publication outputs, either all work should be done within the SLS Safe Setting with other team members coming into review progress/work (or please work with SLS staff to create a synthetic dataset to takeaway to work on from home). Your team members viewing (1) Pre-publication Outputs must have up to date IG training and be listed on your Undertaking Form.
- No cell sizes less than 10, no zeros are allowed unless deemed 'structural zeros' (details below in page 6 on sub-group sizes and statistics)
- **Always provide counts** – ***not*** percentages, proportions, or graphs
- No residual plots

[1] Please note this guide was updated to incorporate info/text from the following Public Health Scotland (PHS) online documents:

- A guide for researchers requesting outputs from the National Safe Haven (PHS) https://www.isdscotland.org/products-and-services/edris/_docs/Guide-for-researchers-requesting-outputs-from-the-NSH-v1-2i.pdf
- Statistical Disclosure Control Protocol (PHS) https://www.publichealthscotland.scot/media/2707/public-health-scotland-statistical-disclosure-control-protocol.pdf
- National Safe Haven: Requesting Outputs (PHS) https://www.isdscotland.org/Products-and-Services/EDRIS/_docs/National-Safe-Haven-Disclosure-Control-v2i.pdf

- If the project has NHS health data (via eDRIS) then all statistics needs to be based on at least 5 sample members (ie not one sample member having repeated admissions)
- Request the tables and produce charts away from the SLS Safe Setting

Your SLS SO will apply SDC checking to clear your Stage (1) Pre-publication output(s) as soon as possible thereafter, and will try to clear within 10 working days. However, Stage (1) Pre-publication Outputs that are lengthy and\or not clearly described can take longer.

Once cleared, Stage (1) Pre-publication cleared outputs will be encrypted and emailed to you. The Stage (1) Pre-publication Outputs (ie non-disclosive outputs) can ***only be shared with those who have signed the SLS Undertaking Form and have valid information governance (IG) training (ie ONS SRT/MRC).***

Your SLS SO cannot clear Stage (2) Publication Outputs. These are cleared by the NRS Data Custodian and you should allow 15 working days.

You should have read the SLS Disclosure Control Protocol for general guidance, available in Annex A.

Stage (1) Pre-publication Outputs Clearance

At the SLS-DSU we understand that it may be useful to report back to your team members (must have signed the SLS Undertaking Form and have valid IG training), we now ask that you please keep any ‘intermediate outputs’ (ie those for internal team discussions) to a minimum.

2020/21 update

All researchers are now offered synthesised versions of prepared/main data files so that all data management through to preliminary statistical analysis can be done on the synthetic extract (see Annex D for guidance on preparing a data extract for synthesis). Meaning there should only be **one** set of pre-publication output requested per project.

For older SLS projects

As noted, we expect that you will keep your Stage (1) Pre-publication Outputs to a minimum. Given at the preliminary stages of projects the sample can change and many variables will be recoded/derived, making it hard to keep track of possible disclosure risks through differencing between tables and sample versions.

The SLS-DSU operates as a service to provide research access to the SLS data. We endeavour to facilitate the process as much as we can but it is important to bear in mind that, in normal circumstances, researchers ***should not expect*** SLS-DSU support staff to provide substantive input into the analysis, research design and interpretation of the results of SLS projects unless they are formally recognised as part of the research project team. As such SLS-DSU support staff time, cannot be allocated to clearing multiple revisions of Stage (1) Pre-publication Output, especially as checking for disclosure risk between revised versions takes many times longer than a simple clearance.

If you are working through output for presentation to your team (e.g. PhD supervisor or project collaborators) which could lead to discussion points or questions as to how to proceed etc, where possible you should request your colleague(s) be present with you in the SLS-DSU Safe Setting (booked into the other PC) to address these directly while looking at the data - in order to produce only the Stage (1) Pre-publication outputs that are required.

For Stage (1) Pre-publication Outputs please minimise the use of exact numbers and percentages. Instead, researchers could use rounding which retains the patterns in the data whilst avoiding disclosure risk. Alternatively, results can be summarised without using numbers/percentages in a narrative style, For example:

- There was a significantly increasing trend in unemployment as deprivation increased
- The table shows how the SLS variables have been re-coded, the smallest group listed here accounted for >20% of the sample i.e. all the re-coded groups are sufficiently large.

Other ways of reporting frequency information or modelling results include producing simple colour coded table cells indicating high, medium and low counts/coefficients/odds ratios should be considered.

However, producing plots/graphs is ***not*** an alternative. It is preferable for SDC checking that researchers take away tables of numbers/stats and produces these outside the safe-setting. **For all graphs and plots submitted for SDC checking, the underlying numbers (n) *must* be supplied** to aid the SLS staff in doing the SDC checking.

Please bear in mind that time spent carefully preparing Stage (1) Pre-publication output in this way is likely to save time in the long-run because it will reduce the time spent by SLS-DSU support staff on SDC checking. In the interests of efficiency, support staff may well prioritise output that makes a clear attempt to assist with the SDC process.

Stage (1) Pre-publication Output clearing – the process

Clearing any Stage (1) Pre-publication Output *may* take **up to** 10 working days during busy periods, especially for more complex and/or lengthy output. When asking for Pre-publication Outputs to be cleared please complete a SLS Stage (1) Pre-publication Outputs Clearance form for each output/group of outputs in one request (see Annex B), and ensure that your outputs comply with all items on the Pre-publication Outputs checklist – it is **your responsibility** to check that your output will pass SDC checking before you submit it.

You may have to undertake SDC measures and save a cut-down, altered or supressed version to be cleared. Examples of SDC measures are given later in this document. Please remember that the output needs to be understandable in terms of what variables (or derived variables) and analyses are being presented, so providing explanatory notes can be helpful. For percentages, graphs and plots please always provide the all the underlying n (sample and category counts). For model output, please provide details of the sample size (model N). Also, please remember to allow time during your safe-setting visit to discuss/describe your outputs to SLS support staff - if they can understand it fully then it will take less time to clear. SLS-DSU support staff can advise on how best to present information/output for clearing if need be.

Please note that if we cannot clear the output as it stands (eg disclosive output needs to be re-worked or output not clearly labelled) then the whole process can take longer than 10 days.

Each Stage (1) Pre-publication output is to be saved as a new folder *To_Be_Cleared* within your '*Reports'* folder along with the corresponding SLS Pre-publication Outputs Clearance form (given in Annex B). Please name each new folder of output successively (1=first output etc.) and date it (SLS staff can show you examples of the file naming we use). Please do not delete this folder after the files have been cleared as they can be useful to refer back to at a later point (especially if substantial SDC was applied). Once SDC checking is completed and the output is cleared, then the cleared output is saved in a separate subfolder *Reports/Cleared* by Support staff (this version is emailed encrypted to you).

Finally, please email your support staff to let them know you have new outputs you would like to be cleared. Support staff do not check the "to be cleared" folders so will not automatically know you have added output and an output clearance form.

Stage (1) Pre-publication Outputs – applying your own SDC checks

In general, please take great care when presenting output to be cleared. It is vitally important that you **do not** present output to be cleared which you know is not permitted. If the outputs cannot be cleared by SLS-DSU support staff because it contains material that presents a disclosure risk, further time and work will be required from the researcher until the output is suitably formatted for clearing. This is particularly important to remember when working to a deadline or in situations where it will be difficult to re-visit the SLS-DSU Safe Setting easily.

Users must also adhere to and be aware of the restriction level of certain variables noted below. Of particular note is level 3; these variables can be used to create derived variables but the raw variables cannot be reported in any analysis. Details given in the data dictionary.

RESTRICTION LEVEL	EXAMPLE VARIABLES	Permissions		
		SLS ADMINISTRATORS	SLS SUPPORT STAFF	SLS EXTERNAL RESEARCHERS(SAFE SETTING ENVIRONMENT)
1	DOBDY JTITLEO INDDES9	Access	No access	No access
2	POSTCODE EASTINGS GRNORTH MIGPCPO CATT	Access	**With permission from the SLS Manager, can:** -View fields -Create new derived variables based on these fields -Provide data to External Researchers based on these fields	No access
3	DATAZONE SIMDSCORE4 CARSCO9 PERSNUM9 DOBMT DOBYR	Access	**Can:** -View fields -Create new derived variables based on these fields -Provide data to External Researchers based on these fields -Link to approved lookup tables via these fields	**Can:** -View fields -Create new derived variables based on these fields -Link to approved lookup tables via these fields **Cannot:** -Remove from the SLS (i.e. report findings on these variables)

Pre-publication and Publication Outputs Clearance Criteria

You should be familiar with SDC, types of disclosure, and measures which can be taken to adjust outputs so that they pass SDC from your Information Governance training. While Annex A details the overarching SLS Disclosure Control Protocol, below are some of the key aspects specific to creating outputs using the SLS. There are also some suggestions of outside sources of further information and guidance.

Analysis and Results

The Public Health Scotland (PHS) SDC Protocol document introduces different kinds of potential disclosure, including Individual and Group Attribute Disclosure, Residual Disclosure (or "Differencing"), Differencing (To produce Small Numbers) and Geographical Differencing[2]. It can be useful to read the PHS SDC Protocol to understand what the different risks are. At the SLS-DSU the main SDC risk – given that the SLS is a 5% sample of the Scottish population - usually occurs is via small numbers (counts) and differencing between tables from previously released outputs (ie a researcher could have previously been looking at retirement ages 65+, but then change the analytical cohort to 66+ which then leads to small numbers differencing between tables when broken down by age-group, sex, etc).

Sub-group sizes and statistics

Many outputs contain frequency counts or cross-tabulations (N values). The guideline is that **N must be ≥10**. If N is 10 or above, it can be reported but if N is between 0 and 9, it cannot be reported. However, in certain cases zeros are allowed if this is to be expected (structural zeroes). For example, in a cross-tabulation of age-group and marriage status we would not expect anyone aged under 16 to be married, so it would be permissible to denote the zero in this case.

Further, we cannot release tables with columns and rows dominated by zeros (or 100% rates), due to the risk of group attribute disclosure.

No hidden columns or rows in tables – and to make clear how any SDC has been applied.

If the project uses health data the count of '10' (events/admissions etc) must relate to more than 5 SLS members (rather than multiple rare events/admissions happening to 1 SLS member). This means that researchers should produce additional tables for clearing purposes when producing any output statistics from health data.

Summary statistics

Maximum or minimum values should only be reported where at least 10 cases share that value. The range should be reported only where both the maximum and the minimum can be reported. When reporting maximum and minimum values in the Stage (1) Pre-publication Output to be cleared, users should in a separate output file provide a frequency count (n) to demonstrate that there are at least 10 cases in the maximum and minimum.

[2] https://www.publichealthscotland.scot/media/2707/public-health-scotland-statistical-disclosure-control-protocol.pdf

It is possible to report percentiles if they are shared by a total of 10 observations. Similarly, the mode should only be reported if there are at least 10 observations at the modal value.

For numerical data with many unique values it may be better to report the percentage of values in certain fixed ranges: e.g. income levels in certain bands. The mean, variance and higher order parametric statistics can be reported if N is 10 or above.

Where possible, present means, medians, inter-quartile ranges and standard deviations.

Graphical output

It is preferred that charts/graphs/plots are produced at a later stage away from the SLS-DSU Safe Setting, so tables of numerical values to produce these can be requested as output instead. If charts/graphs/plots are produced in the Safe Setting (in cases where these are a by-product of a statistical method within in a stats package which, for example, also generates CIs etc), the corresponding tables of the underlying data (n) ***must*** also be produced for clearing, this is to demonstrate that there are no counts/cell sizes <10.

Histograms and bar charts (whether simple or stacked) will not pass SDC if the same data in the frequency table form is not also reported. The frequency table used to produce these charts ***must*** also be provided for clearing. The use of scatterplots is discouraged, however if the data plotted are derived from models then this may be permissible and should be discussed with SLS Support Staff.

Generally, for graphs/charts/plots to pass SDC, it ***must*** meet the following criteria:

- additional tables with the counts are produced alongside to demonstration that no bars/points represent <10 SLS members/sample members
- data points cannot be identified with units (when the graph consists of smoothed data this is not usually a problem)
- there are no bars/points where the underlying data represents <10 SLS members/sample members
- there are no outliers that might lead to the identification
- the graph is submitted as a fixed picture, with no data attached. This means graphs should be image files (either .jpg, .jpeg, .bmp or .wmf – *not Stata charts*).
- Kaplan-Meier survival plots should be smoothed (SLS-DSU support staff can help advise on deaths as the rules are slightly different).

Residual plots ***should not*** be requested, please discuss with SLS-DSU support staff.

Tabular output

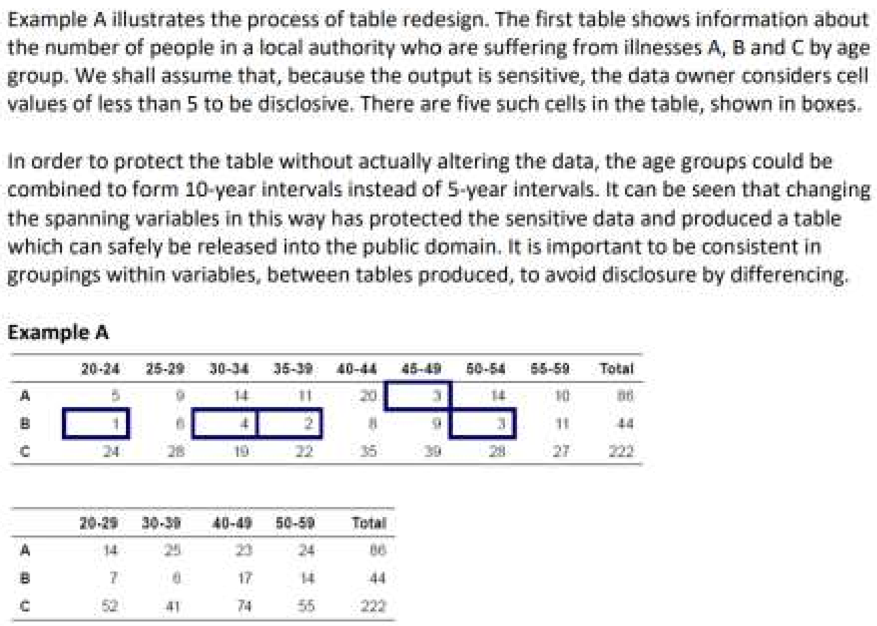
As noted, these should only be requested as Stage (1) Pre-publication Output to produce Stage (2) Publication Outputs. As the SLS is a 5% sample, producing tables for publication often requires SDC measures to be used. There are various ways to address dealing with small numbers. These are detailed below and further examples can be found in the PHS SDC guide[3]. Remember **no cell counts <10**.

Suggested SDC Measures

Table Redesign

This is the preferred SDC option at the SLS-DSU as it ensures that there are large enough cell sizes and usually minimises the risk of residual disclosure (or "differencing") between subsequent outputs.

The PHS example below explains the issue and proposes a solution:



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

	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	Total
A	5	9	14	11	20	3	14	10	86
B	1	6	4	2	8	9	3	11	44
C	24	28	19	22	35	39	28	27	222

	20-29	30-39	40-49	50-59	Total
A	14	25	23	24	86
B	7	6	17	14	44
C	52	41	74	55	222

Source: Statistical Disclosure Control Protocol (PHS)[4]

Suppression within tables

We advise users to group categories or adapt output via 'table redesign' to avoid the need for suppression in tables, however, in certain circumstances, suppression may have to be used.

It is assumed that row and column totals are reported as part of the table or, alternatively, that they can be calculated from outputs published elsewhere from the same data set.[5] Not displaying the row and column totals **cannot** be used as a method of SDC. If your results are being presented as percentages you ***must*** supply the numbers as additional material to assist

[3] www.publichealthscotland.scot/media/2707/public-health-scotland-statistical-disclosure-control-protocol.pdf

[4] www.publichealthscotland.scot/media/2707/public-health-scotland-statistical-disclosure-control-protocol.pdf

[5] Not to make this assumption would involve keeping a record of the marginal totals in question and ensuring that they are never published. As this would in time become unmanageable, the assumption that marginal totals will be published should be made from the start.

clearing – we would prefer only tables with counts (then percentages/proportions can be created later away from the Safe Setting).

If a single cell is suppressed, it will be necessary to either:

(i) suppress one or more cells in the same row **and** in the same column until the totals of the suppressed cells in both the row and the column are at least 10

(ii) to merge the cell with an adjacent one in the same row or column until the merged cell frequency is at least 10.

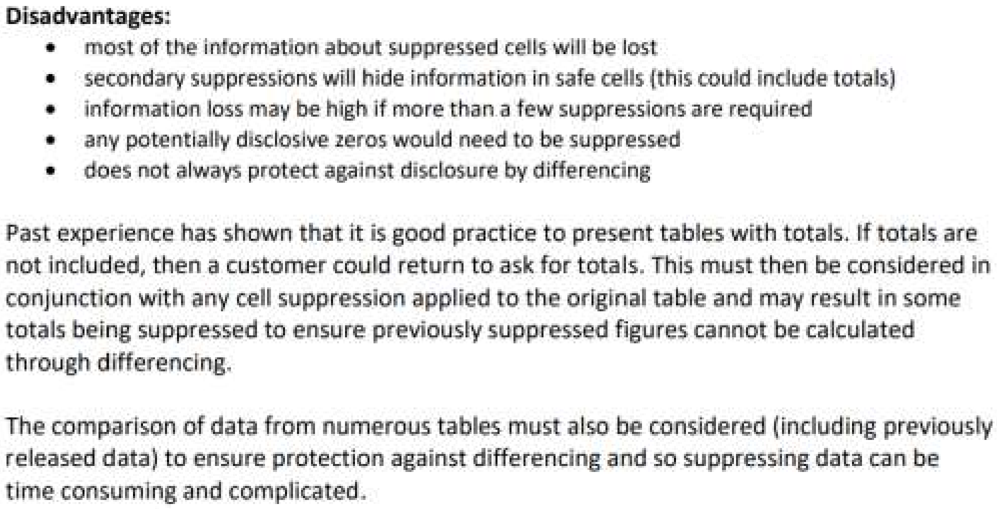
This latter option of combing categories would be preferred, depending on what is being displayed this need not involve merging the entire row/column with its neighbour. The "merge cells" option on Excel or Word can be used to minimise the number of cells to be merged. The symbol '.' should not be used – suppression should be indicated by ***.

For the toy example below, 4 cells must be suppressed here to avoid disclosure that only 1 case is contained in cell "BX"[6]:

Table as received			
	X	Y	Total
A	17	17	**34**
B	1	16	**17**
C	15	15	**30**
Total	***33***	***48***	***81***

Table with suppressed cells			
	X	Y	*Total*
A	17	17	***34***
B	***	***	***17***
C	***	***	***30***
	33	***48***	***81***

PHS summarises the disadvantages of cell suppression within tables:



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.

Source: Statistical Disclosure Control Protocol (PHS)[7]

[6] In practice we would prefer combining categories via 'table redesign' rather than cell suppression as in this simple toy example using the marginal total and iterative proportional fitting (IPF) these supressed cells could be reidentified. More on IPF from: https://eprints.whiterose.ac.uk/5029/5/IPF-Norman-SoG-WP99-03.pdf

[7] www.publichealthscotland.scot/media/2707/public-health-scotland-statistical-disclosure-control-protocol.pdf

Rounding small cells in tables

This approach is applying controlled rounding to a multiple of a set base such as round to the nearest 0/10 or 0/50. While this method is routinely used by SLS researchers, rounding to an arbitrary value may reduce the usefulness of the data being presented. Note that care must be taken when requesting further tables without rounding applied in order to avoid the risk of residual disclosure (or “differencing”) between subsequent outputs.

Cell swapping within tables

This approach is applying targeted or random cell swapping methods. For example, if one category has a cell size of just 9 in order to be allowed for SDC release to ‘take’ a 1 from a surrounding cell and then to work back all the table margins to align with the new cell of 10 and the corresponding cell that has lost 1. This method is used by SLS researchers, swapping with targeted cells to meet SDC requirements, however, this approach may reduce the usefulness of the data being presented as its changing more than one attribute of the data being presented. Note that care must be taken when requesting further tables without cell swapping applied in order to avoid the risk of residual disclosure (or “differencing”) between subsequent outputs.

Additionally, please provide clear information on how the cell swapping was done – ie to leave the raw unswapped table, any intermediate stages and the resulting ‘swapped table’ to allow SLS-DSU staff to fully understand what SDC has taken place during SDC checking.

Impact of SDC on tables - disclosure between tables

Codings of variables should wherever possible be kept consistent. If age is to be coded in bands of X years, different Stage (1) Pre-publication output tables should not have different codings as this may allow a disclosive frequency to be calculated by subtracting one table from another. For example, if one table used an age band of 5-9 years and another table used an age band of 5-10 years, then subtracting would show the data for those of age 10 years.

In the unusual situation that there is a good research reason that multiple codings are to be used, you must ensure that any cell count derived from the differencing between tables is over 10. This applies to all tables produced throughout the entirety of the project.

To avoid this, please delay wherever possible, requesting any Stage (1) Pre-publication output of numbers until you have determined what your final sample/groupings are.

Further, if a statistic is suppressed in one table it must not be derivable from data released in any other table. As an SLS user it is your responsibility to ensure that this does not occur. If this does occur and there have been no Stage (2) Publication Outputs produced, to have the later table cleared you will be asked to confirm deletion of the earlier Stage (1) pre-publication output table before you are permitted to receive the later table. If the first table has been presented as a (final) Stage (2) Publication Output then the later table cannot be cleared.

Statistical Models

More straightforward from a SDC perspective; however, some things to be addressed:

- All models Ns (ie the number of cases a model is based on) to be provided.
- Sufficient observations for the model or at least 10 residual degrees of freedom as appropriate.
- If warnings given due to 'perfectly predicts failure' this could be disclosive on the category from the variable in the model (ie everyone age 90+ has the outcome being modelled, but say there are only 6 sample members 90+ group then it is essentially a count). As such, a frequency count of what was flagged from the modelling must be provided.
- If adding in age in years when setting time or for multilevel modelling using the sample member as the level, information/text from the model set up should be removed as SDC as not to detail the oldest SLS sample member or how many times a sample member has the 'event' being modelled (ie hospital admission etc).

We cannot emphasise enough the importance of ensuring your output meets the criteria set out in Annex A. If your output does not meet this criteria you will be required to attend the SLS-DSU Safe Setting in person to discuss with SLS-DSU support staff and rectify the output before the output can be released to you[8].

General Guidelines for Stage (2) Publication Outputs Clearance

When you want to disseminate your SLS results beyond your project team and named associates in the SLS Undertaking Form you **must obtain Publications outputs clearance** from the SLS Data Custodian using the SLS Publication Clearance Request Form (Available from the SLS website step 12). If a member of your research team produces a Stage (2) Publication Output (previously referred to as Final Outputs), it is the responsibility of the approved researcher to ensure that the publication output goes through the Stage (2) Publication Outputs Clearance process.

Typical Stage (2) Publication Outputs include conference abstracts, presentations for department seminars through to conference talks, working papers, reports or journal articles intended for publication. The SLS Data Custodian must clear all types of Stage (2) Publication Output and you should **allow 15 working days** for publication outputs clearance (though keep in mind it may take up to 20 days during busy periods). Although most Stage (2) Publication Outputs are cleared more quickly than this, larger outputs such as PhD theses will take longer and you should build into your PhD submission timetable enough time for the possibility of the SLS Data Custodian asking for changes before giving approval. The process for clearing Stage (2) Publication Outputs reduces the risk of disclosure, ensures that the study and data are properly described and that the data have been used appropriately.

Key criteria that will be considered are:

- The results displayed and the discussion concerning them do not raise confidentiality or disclosure issues;
- The SLS is described correctly;

[8] For overseas researchers you will be required to Skype/Teams/Zoom call with SLS-DSU support staff at a time convenient for them.

- 'Source: Scottish Longitudinal Study' is added to tables and figures, where appropriate;
- There is no reputational damage to the Scottish Government, The National Records of Scotland and the Scottish Longitudinal Study
- You have acknowledged the support of the LSCS using this disclaimer:

"The help provided by staff of the Longitudinal Studies Centre - Scotland (LSCS) is acknowledged. The LSCS is supported by the ESRC, National Records of Scotland and the Scottish Government. The authors alone are responsible for the interpretation of the data. Census output is Crown copyright and is reproduced with the permission of the Controller of HMSO and the King's Printer for Scotland."

Please keep in mind that the Stage (2) Publication Outputs Clearance checks should not be considered a peer review process as the NRS Data Custodian will only check the key criteria as described above.

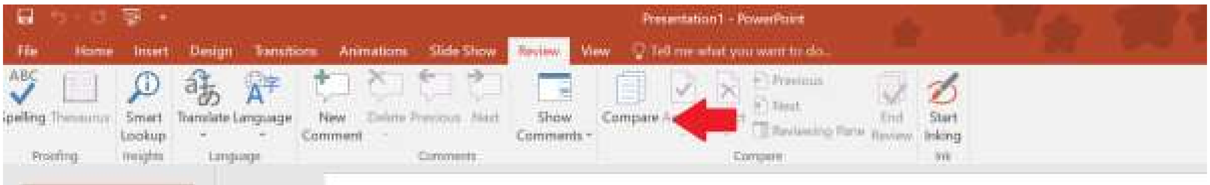
Errors in the Interpretation of the data or in the analysis are entirely the responsibility of the researcher.

When you submit a Stage (2) Publication Output Clearance Form to you will also submit a completed checklist (see Annex C). Similar to the Stage (1) Pre-publication Output process, you will normally be sent an Outlook calendar invite so you have a record of the (latest) date by which you should expect to receive your clearance.

It is therefore essential that you plan to allow enough time at the end of the day to thoroughly check your output to avoid potential delays and to allow enough time for the SLS-DSU support staff to clear the Stage (1) Pre-publication Output and then the NRS Data Custodian to clear the Stage (2) Publication Output if you have a tight publication deadline coming up. If in doubt, please discuss with SLS-DSU support staff.

Reuse of Publication Output – primarily for talks:

1. ***If the output remains unchanged*** and there is no new discussion of SLS processes, once your Stage (2) Publication Output has been cleared for public release you may disseminate it a number of times without resubmitting it for clearance. However, if the output text or results changes at all then it will be required to be re-cleared. We would kindly remind you that we do require for ESRC reporting purposes information on where and when your research using the SLS is presented. This extends to future unchanged presentations which do not need clearing.
2. ***Any changes in tables, figures, text or content*** requires the submission of another Stage (2) Publication Outputs Clearance Form. For example, if you need to submit to a different journal then you should submit a Stage (2) Publication Outputs Clearance Form for each submission.
   For speed of clearance, you can use 'track changes' if your output is in 'Word' format or 'compare and combine' for outputs in 'Powerpoint' (see image below) to ensure SLS-DSU support staff can quickly see the changes.



If you do not do this, clearance may be delayed as the NRS Data Custodian will need to check for changes manually.

As you have gained either SLS Approved Researcher status or Provisionally Approved Researcher status, the SLS-DSU trust that you will inform us of any changes and not release/present/publish any output which has not been officially cleared especially that which may inadvertently cause reputational damage to the SLS-DSU/NRS/Scottish Government. Releasing uncleared data would be a breach of the terms and conditions of your Undertaking with the SLS.

Again to kindly stress you **MUST** notify support staff (or the SLS-DSU on sls@lscs.ac.uk) when any publication which draws on the SLS is published. We maintain a database of all published research outputs that use the SLS and it is vital for us to keep this record up-to-date (https://sls.lscs.ac.uk/outputs/)

In all cases, the Registrar General reserves the right to withhold clearance of any output if such an output would inhibit the Registrar General's ability to carry out their statutory duties.

Stage (2) Publication Outputs Clearance Criteria

Describing the SLS (see LSCS Working Paper 1.0 for further details)[9]:

The SLS must be described correctly ensuring accuracy of the text and of substantive points about the SLS and its functions, for example:

- the SLS should not be described as being used to "track people";
- the SLS should be referred to as a database: data provided for researchers to use in analysis should normally be referred to as "datasets";
- the SLS linkage method, sample size and content (i.e. that it includes data for Scotland only) and study methodology (see below) should be described accurately;

Researchers must make clear that SLS linked datasets have no identifiable individual level data and are derived from linkages that are anonymised prior to handover to the research team. For example, if your support staff linked your datasets using the postcode (deleting the postcode before providing you with the data) then you must state that SLS-DSU staff carried out the linkage and not say that 'we' linked the data.

[9] http://calls.ac.uk/wp-content/uploads/2013/05/LSCS-WP-1.0.pdf

Methodology

- The SLS sample is made up of 5.3% of the Scottish population
- The SLS sample is selected using 20 dates of birth.
- Values cannot be described as "missing", it is more appropriate to refer to them as "non-response (missing/edited)".

Please note that 2011 data have imputed values and no missing values. Imputed values occur when the original data was either missing or has been edited. 2011 data before imputation are not available so this difference between 2011 and 2001/1991 census data should be considered when interpreting results. For example, it is not true to say that the proportion missing has decreased in the 2011 census. We have some imputation flags for primary census variables such as age, sex i.e. responses to single census questions (see the data dictionary) however some SLS variables are derived from several primary or secondary etc. variables and the degree of imputation becomes very difficult to work out. If you have requested any imputation flags as part of your project variables you are **not** allowed to publish any tables or report any specific statistics based on analysis using the imputation flags. The imputation flags are for informational purposes only. You can use them to refine samples etc.

Analysis and Results

The same rules that apply to the level of output control for the named researchers are also applied to information that is to be released publicly using SLS data.

In general if N is between 0 and 10, it cannot be reported, although in certain cases (e.g. reporting on rare diseases, deaths) it may be permissible to report on N-values between 5 and 10, however this will be at the discretion of the SLS-DSU and NRS.

Thus, the minimum N of 10 also applies to tables and summary statistics (ie if reporting max and min frequency tables of these variable should also be provided).

Annex to the SLS Undertaking Form

Immediate sanctions to be applied in the case of breach of the conditions of the undertaking

All users of the Scottish Longitudinal Study (SLS) who have signed the SLS Undertaking Form must report any breach of the conditions of the Undertaking promptly to the NRS SLS Project Manager. Failure to do so is a fundamental breach of the terms and conditions of the Undertaking. It should be noted that in signing the Undertaking individuals (and their institutions) are agreeing to the terms of the Census Act (1920), the Statistics Act (1938) and current Data Protection and Freedom of Information legislation.

The following sanctions may be applied:

> Step 1: For a first offence, depending on the seriousness, the penalty should be a minimum 1 month discretionary suspension from access to any SLS or NRS data applicable to the individual(s) in question. It would (i) generate a written warning to that individual's institution of employment and (ii) require ONS SRT training to be redone. A more serious breach will handled as a Step 2 breach. Subsequent offences would be escalated to Step 3.
>
> For a more serious breach, a minimum of 12 months discretionary suspension of access would be applied, as detailed in step 2. Subsequent offences would be escalated directly to step 3. This would also be applicable to other researchers named on that individual's SLS project(s).
>
> Step 2: For a second offence, the penalty should be a minimum 12 month discretionary suspension from access to any SLS or NRS data applicable to the individual(s) in question. This would also be applicable to other researchers named on the SLS project. Again, it would generate a written warning to that individual's institution of employment.
>
> Step 3: An individual's further breach would, as a minimum, result in a suspension of access of 2 to 5 years, or permanently, on the individual and would generate a written warning from the Responsible Statistician (the Registrar General for Scotland).
>
> Or where the breach is the result of an institution's wilful or negligent action, then a minimum penalty of a 12 month non-discretionary suspension shall apply to the relevant department within the institution. Repeated breaches will result in a letter from the Responsible Statistician with discretionary penalties to the institution as a whole, including suspension of all SLS and NRS data access facilities for all the institutions staff.

Please also refer to the NRS 2022 Census Confidentiality Undertaking (CCU) which you signed and emailed to the NRS SLS PM at the start of your project.

Further information on how NRS protects the confidentiality of census data from: www.scotlandscensus.gov.uk/confidentiality and www.scotlandscensus.gov.uk/Privacy-2021

Annex A

Scottish Longitudinal Study Disclosure Control Protocol

This document must be read by all persons wishing to use SLS data

The SLS is a linked database containing individual confidential data, managed by the Longitudinal Studies Centre – Scotland (LSCS). Any person using that data, whether in the role of a SLS-DSU support staff providing data to a user, or as an end user receiving results of data for research, must comply with the confidentiality requirements stated in the undertaking.

Further to this, certain disclosure controls must be applied to the data to ensure that no individual can be identified within them. These controls will be applied to any data released to users by support personnel. If data are to be released in tabular form then the SLS-DSU support staff must ensure that any variables that alone, or in conjunction with others, may identify individuals are aggregated to the point where no identification is possible.

- No data on the birth dates of SLS members may be released, with the exception of year of birth. Where full date of birth is required for use in derivations (i.e. in such procedures as person years at risk analysis) only those SLS staff based at NRS with full database permissions will be allowed access to the data.

- Exposure times (e.g. person years at risk) may be included in aggregated datasets provided ***either*** there is more than 10 events in each cell ***or*** else the data has been subject to adjustment to prevent disclosure.

When releasing tabular data SLS support staff must ensure that cell counts are 10 or over for Pre-publication and Publication Outputs. If associated data allows the cell to be split then the support person must aggregate the data to the highest level consistent with the need to explain the results.

- Sample uniques are never allowed.

- Reporting residual values that identify individual cases will not be allowed when releasing data from statistical models.

- Particular care should be exercised with plots especially where there are outliers or extreme values. These should not be released.

- Histograms and bar charts can be reported if, and only if, the same data in frequency table form could be reported.

Restrictions will also be placed on the release of any variable deemed to be sensitive. These include variables which relate to small numbers of people in Scotland (i.e. local-area geographic identifiers, detailed ethnicity, rare causes of death etc.). Other variables, such as religion, may also be treated as sensitive, depending on the context of the research. It should be noted that selection criteria used in extracting data such as sex and age may be disclosive when used in conjunction with other variables.

If a SLS-DSU support staff believes that data may be disclosive they must bring this to the attention of the NRS SLS Data Custodian who will decide on the procedure to follow. In most cases this will require further aggregation of the data.

Annex B Pre-publication Outputs Clearance Criteria

SLS Stage (1) Pre-publication Outputs Clearance Form[10]

Project Number: ____________________________

Researcher Name: ___________________________________

Date of request: ______________________________

Folder location and name:
Reports\To_be_cleared\____
(eg \Reports\To_be_cleared\2018_tbc(n)_yymmdd) (where n=output number, yymmdd=date)

Date of SLS approval (SLS use only) ______________

Approved by (Staff initials - SLS use only) ______________

I have read the SDC guidance v3.5 document ☐

I have read the Pre-publication Output Clearance Guidance and completed the Statistical Disclosure Control Checklist (below) ☐

I confirm that the requested outputs fall within the scope of the project's aims and objectives ☐

Please complete the table below for each file to be released (insert additional rows if required)

File name	Description of file contents	Is this an update to a previously released file? If so, please provide details of changes

[10] Please note this SLS form was adapted from the eDRIS doc: "*National Safe Haven Statistical Disclosure V3*"

Stage (1) Pre-publication Output Clearance GUIDANCE

As far as possible, limit pre-publication output requests to one of the following scenarios:

1. Pre-publication i.e. necessary to produce your (final) Stage (2) Publication outputs
2. Discussions with team members, without which the analysis would be delayed. Team members should attend the Safe Setting where possible and researchers should make the majority of analysis decisions themselves

Check the following prior to requesting a file to be released from the SLS (see also the more detailed disclosure control checklists on p. 3-4):

- Titles and descriptions are clear and self-explanatory. Tables, charts[11], variables and variable codes should be labelled in a meaningful way and have been formatted to a standard suitable for a final presentation i.e. not cut and pasted from log/output windows
- Outputs do not contain embedded data which could be made available after release
- There are no small, potentially-disclosive numbers - no tables should have a cell count of less than 10
- The sample used in each table/chart/model is shown i.e. description including exclusions and sample size (N). Any statistics should be based on a sample size of at least 10
- Differencing from previously released outputs: Have you produced similar analysis before that, combined with this output, could be used to identify someone?
- Every effort has been made to convert data results to descriptive text to avoid release of actual results. For example, a table of age in 1991 by age in 2001 would not be released due to small numbers but text could be written '*although in most cases age in 2001 is a corresponding number of years older than age in 1991, there are instances where this is not the case so a decision needs to be made to either exclude these cases or accept the 1991 age is being 'correct'. Excluding such cases would reduce the sample size by <5%*'. This approach of summarising results also makes team discussions quicker and easier
- Any graph/chart/percentage must also include the data behind the graph/ chart/ percentage, have no outliers and be in a suitable format. Scatter plots are not released.

Save this completed Pre-publication Outputs checklist form with the outputs and email your SO to request clearance together with the location of the output

If output does not comply with the above, SLS support staff will cease output checking and you will need to return to the SLS Safe Setting to amend the output

[11] Where possible please produce charts from tables away from the Safe Setting as charts should only be produced if the counts can be released

SDC CHECKLIST for SLS Stage (1) Pre-publication Outputs[12]

Complete the following SDC checks: (Please answer all questions)

Frequency tables/charts

	Yes	No	N/A
Are there any cells in the table with a value >0 and <10?	□	□	
Are there any columns or rows dominated by zeros or 100% of observations?	□	□	
Are there any cells with suppressed values/hidden columns or rows?	□	□	
Has the table used a different population base from previous similar tables?	□	□	
Has the table used a different variable breakdown from a previous similar table?	□	□	
Has the table used a different definition or source for a variable previously tabulated?	□	□	
Are there any minima/maxima present? (for max/min separate frequency counts/univariate distribution table must be provided – ie for age could be disclosive for top age or certain 'events')	□	□	
For health data linkage projects			
If using health data does the count of '10' (events/admissions etc) relate to more than 10 SLS members (rather than a rare event to 1 SLS member)	□	□	□

Models

	Yes	No
Does the model have fewer than 10 residual degrees of freedom?	□	□
Does the model description quote or plot any individual values, such as minimum or maximum values or outliers? Again, plots are discouraged and tables to be produced to produce away from the Safe Settings	□	□
Does the model description include a residual plot or residual values?	□	□
Has the model used a different population base (ie cohort) from a previously described (or cleared) model (or sets of models)?	□	□
Is the regression undertaken on a single unit?	□	□
Does the regression solely consist of categorical variables?	□	□

[12] Please note this SLS form was adapted from the eDRIS doc: "*National Safe Haven Statistical Disclosure V3*"

Syntax files

	Yes	No
Is the code clearly annotated with comments to assist the reviewer?	▨	□
Are there any references or figures in the comments or code that could lead to potential identification of individuals?	□	▨
Are there any SLS numbers included in the code or the comments?	□	▨
Are there any counts from the data present in the comments or code?	□	▨

- **Any white** boxes ticked, your output will fail SDC. If you think your output should still pass, please discuss this with your SLS support staff before leaving the safe setting

- **Any light grey** boxes ticked, your output may fail SDC. If your output fails, your SLS support staff can also provide advice about how to re-design your outputs so they will pass

- **All dark grey** boxes ticked, your output is likely to pass SDC, but it still needs to be checked

The SLS support staff will compare your outputs to all outputs previously released from the same dataset to make sure it is not possible to identify any individuals or small cells by comparing outputs.

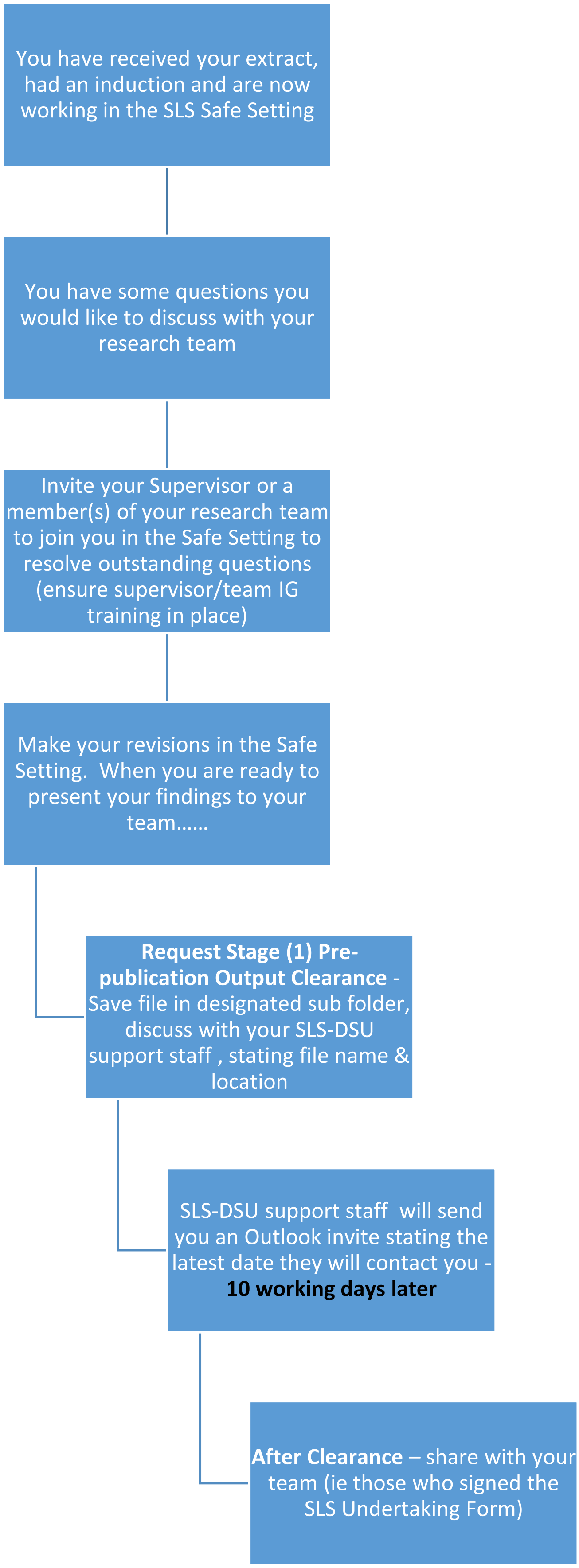
Annex C Stage (2) Publication Outputs Clearance Criteria

CHECKLIST Stage (2) Publication Outputs Clearance Criteria

- The SLS is described correctly □
- 'Source: Scottish Longitudinal Study' is added to tables and figures, where appropriate □
- The support of the LSCS is acknowledged using the specified disclaimer □
- No tables have a cell count of less than 10 (unless previously agreed with SLS-DSU support staff) □
- There is no disclosure between tables □
- Any statistics should be based on a sample size of at least 10 and when reported should include confidence intervals and/or statistical significance □
- Any graphs should be based on cleared output, have no outliers and be in a suitable format □
- There is no reputational damage to the National Records of Scotland, The Scottish Government and the Scottish Longitudinal Study. □

□**I have read and understood the Statistical Disclosure Protocol (Annex A)**

Process chart for Stage (1) Pre-publication output clearance



You have received your extract, had an induction and are now working in the SLS Safe Setting

You have some questions you would like to discuss with your research team

Invite your Supervisor or a member(s) of your research team to join you in the Safe Setting to resolve outstanding questions (ensure supervisor/team IG training in place)

Make your revisions in the Safe Setting. When you are ready to present your findings to your team……

Request Stage (1) Pre-publication Output Clearance - Save file in designated sub folder, discuss with your SLS-DSU support staff , stating file name & location

SLS-DSU support staff will send you an Outlook invite stating the latest date they will contact you - **10 working days later**

After Clearance – share with your team (ie those who signed the SLS Undertaking Form)

Annex D: Guidance for researchers on synthetic data

SLS-DSU

Scottish Longitudinal Study
Development & Support Unit

This document is to provide guidance for researchers who wish to have a Synthetic version of their main SLS dataset to take away from the SLS Safe Setting

Using the R software package 'synthpop' the SLS-DSU team have the ability to produce non-disclosive versions of your main SLS analysis dataset that you can then take away to work on outside the safe haven on your institutional device (however, this data must **never** be passed onto another party). A SLS Synthetic Data Undertaking form must be signed to confirm this – this also means NOT passing on to project team members, unless they have also signed the SLS Synthetic Data Undertaking form.

Because this synthetic data is model-based, it retains some of the statistical structure of the 'real' data and therefore you will often find that models run on the two datasets are very similar. You can use your synthetic data to develop your methods, but you should ***always*** base your research on the actual data. As detailed in the SLS synthetic data SLS Synthetic Data Undertaking form you must **never** present or publish results based on the synthetic data.

In order to produce synthetic data we need you to pre-process your data. This should not be extra work because in most cases the restructured dataset will be the same or close to the one you would want to carry out a 'final analysis' on (ie after your main data management).

Preparation of your data for synthesis:

1. It must be in **wide format** i.e. a simple flat file with a single case in a row of data. At an initial stage of your project you could have more than one of these and you can get each one synthesised, but it will not be possible to link the synthesised files.

2. Remove all identification numbers (usually **sls number**).

3. Replace any date variables by numeric or categorical variables, e.g. use 'age at death' and/or 'follow up time' instead of 'date of death'.

4. Identify the following to SLS support staff:
    a. Any continuous variables with missing values that are indicated by something other than the system missing value (e.g. -9 or 999). System missing values are fine if you have only one missing value code (e.g. in STATA).

   b. Optionally, one grouping variable of major interest that divides your data into large groups of at least 500 records each (e.g. Local Authority, or Health Board). This will be used for stratification.
   c. If possible, a list of at most 5 variables that are your major focus of interest (e.g. your outcome variable and those you expect to be the most important predictors).

5. All categorical or ordinal variables (except the stratification variable) should have a **maximum of 15 categories**. Here are some strategies to deal with this:
   a. Group categories with small numbers of records together
   b. If you have a variable with lots of categories (e.g. detailed cause of death), please provide an additional variable that groups it into fewer categories. We will use this second variable in the synthesis. A synthesised version of your detailed variable will appear in the synthetic data but its relationship with other variables will only be via its grouping variable.

6. Ideally you would provide this data as an R data file (.Rdata or .RDS extension). All variables should be numeric, factor, character or logical (**no other data types**) and stored in a data frame. Alternatively, you can provide a CSV file with the first row including your variable names. If your data are in STATA or SPSS you should speak to SLS support staff about supplying it in this format, as there may be issues with variable label codes being changed in the synthetic version. As such it may be worth asking SLS support staff for the original syntax used to produce your SLS extract (this would be in STATA or SPSS format). Plus, can be useful to request your data management syntax/code created to be cleared to take away as from these 2 sets of code (ie the original SLS labels and your own code) categories can be identified.

7. Outliers, on a continuous variable, must be top and bottom coded. For your ease please simply top code at the 95th percentile and bottom code at the 5th (eg if the 95th percentile for an income variable is £60,000 all incomes above £60,000 should be given the value £60,000). This is most likely to apply to age, hours worked, hours caring etc.

Please read these instructions very carefully and follow them precisely, otherwise valuable SLS support staff time might be wasted. Given that we are a small team and continue with low density working in the SLS-DSU, efficient use of time and resources is crucial.

We therefore reserve the right not to produce synthetic data for anyone who fails to meet these instructions.