SLS Outputs Guidance Document

General Guidelines for Intermediate Outputs Clearance

Intermediate outputs are those which once cleared can be shared with ***only those who have signed the SLS undertaking form.*** You should have read the Scottish Longitudinal Study Disclosure Control Protocol for general guidance, available in Annex A or online at: Disclosure Protocol[1]

Intermediate outputs

While at the SLS we understand that it may be useful to report back to your team members who have signed the undertaking form. However, as much as possible, we expect that you will keep your intermediate outputs to a minimum. At the preliminary stages of projects the sample can change and many variables will be recoded making it hard to keep track of possible disclosure risks through differencing between tables. The SLS operates as a service to provide 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*** support officers to provide substantive input into the analysis, research design and interpretation of the results of SLS projects unless they are formally part of the research project team. Support Officers' (SO's) time cannot be allocated to clearing multiple revisions of intermediate 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 collaborator) and have questions, where possible, you should request they be present with you in the Safe Setting in order to produce only the intermediate outputs that are required.

Exact numbers and percentages should not be required before final outputs – therefore either rounding can be used which will retain the key message(s) from the analysis whilst avoiding disclosure risk or results can be summarised without using numbers/percentages. Examples of more narrative styles of results reporting might include:

- 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 colour coded table cells indicating high, medium and low counts/coefficients/odds ratios should be considered. ***Please bear in mind that time spent carefully preparing output in this way is likely to save time in the long-run because it will reduce the time spent by support officers on disclosure control checking. In the interests of efficiency, support officers may well***

[1] http://sls.lscs.ac.uk/wp-content/uploads/Scottish-Longitudinal-Study-Disclosure-Control-Protocol-2015.pdf

prioritise output that makes a clear attempt to assist with the disclosure risk checking process, so please bear this in mind.

Intermediate Output clearing process

Clearing output may take **up to** 10 working days during busy periods, especially for more complex and/or lengthy output. When asking for outputs to be cleared, ensure that the outputs comply with all items on the intermediate outputs checklist and save them in a separate subfolder in the *Reports* folder. You may have to undertake statistical disclosure control (SDC) measures and save a cut down or supressed version to be cleared. For example in a cross-tabulation with small cells you can either remove the categories with the small counts from the table (ensuring that row and column totals do not reveal the removed categories) or can combine the small counts into larger groups. It is permissible to note below the table that there were small counts in certain categories, but you cannot reveal what the small counts were. The SO will talk through the file naming system for outputs which should include the project number, successive numbering of outputs (1=first output etc.) and date. Do not delete this folder after the files have been cleared. Make sure there is sufficient time to discuss the outputs with the SO before leaving the Safe Setting. This ensures that the SO can understand what recoding of variables may have occurred and the output which they may not be familiar with. The SO should be emailed with a request to clear intermediate outputs including the folder name and the checklist (see appendix 1). They will then send you an outlook calendar invite so you have a record of the date by which you should expect to receive your cleared outputs.

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 the support officer 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 safe settings 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 can be accessed in the data dictionary.

RESTRICTION LEVEL	EXAMPLE VARIABLES	Permissions		
		SLS ADMINISTRATORS	SLS SOS	SLS EXTERNAL RESEARCHERS(SAFE SETTING ENVIRONMENT)
1	DOBDY LINKTYPO ZQUERY 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)

Intermediate and Final Outputs Clearance Criteria

Analysis and Results

Sub-group sizes and statistics

Many outputs contain frequency counts or cross-tabulations (N values). The guideline is that if N is 10 or above, it can be reported but if N is between 0 and 9, it cannot be reported. In general N-values of zero 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 we would not expect anyone aged under 16 to be married, so it would be permissible to denote the zero in this case.

Suppression within tables

We advise users to group categories or adapt output 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.[2] 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 your SO with the numbers as additional material to assist clearing. 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 or (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 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 example, 4 cells must be suppressed here to avoid disclosure that only 1 case is contained in cell "BX":

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***

Disclosure between tables

Codings of variables should wherever possible be kept consistent. If age is to be coded in bands of k years, different 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 the two tables would show the data for those of age 10 years. In the unusual situation that there is a good reason that multiple codings are to be used, you must ensure that any cell count which can be 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, intermediate output of numbers until you have determined what your final sample/groupings are.***

[2] 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.

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 final outputs, to have the later table cleared you will be asked to confirm deletion of the earlier table before you are permitted to receive the later table. If the first table has been presented as a final output then the later table cannot be cleared.

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 intermediate output to be cleared, users should in a separate output file provide the SLS SO with a frequency count to demonstrate that there are at least 10 cases in the maximum and minimum. 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.

Graphical output

Histograms and bar charts (whether simple or stacked) can be reported if, and only if, the same data in frequency table form could be reported. The frequency table used to produce the chart must be provided to the SLS SO.

In general, graphical output can be used only to make tabular or other formats more accessible to the user. 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 your SO. If these are both true then the graph should then meet the following conditions:

- data points cannot be identified with units (when the graph consists of transformed or fitted data this is not usually a problem)
- there are no outliers that might lead to the identification of a unit
- 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).

Kaplan-Meier survival plots should be smoothed.

General Guidelines for Final Outputs Clearance

When you want to disseminate your SLS results beyond your project team and named associates you **must obtain final outputs clearance** from the SLS Data Custodian using the SLS Clearance Form (Available from the SLS website) If a member of your research team produces a final output, it is the responsibility of the approved researcher to ensure that the final output goes through the complete Final Outputs Clearance process.

Typical final outputs will be working papers, reports or journal articles intended for publication, presentations or abstracts. The SLS Data Custodian must clear all types of output and you should **allow 15 working days** for final outputs clearance (though keep in mind it may take up to 20 days during busy periods). Although many 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 final 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;
- '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 Queen's Printer for Scotland."*

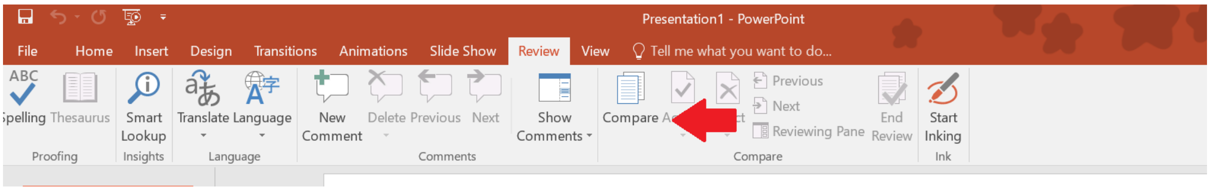
Please keep in mind that the Final Outputs Clearance checks should not be considered a peer review process as the SLS/Data Custodian will only check the key criteria as described above. ***Errors in the Interpretation of the data or in the analysis are entirely the responsibility of the researcher.***

When you submit a Final Output Clearance Form to your SO you should also submit a completed check list (see Annex C). The SO will send you an outlook calendar invite so you have a record of the date by which you should expect to receive your clearance. 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 in person to discuss and rectify the output with your SO before the output can be released to you[3]. 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 Team and the Data Custodian to clear the output. If in doubt, please discuss with your SO.

If the output remains unchanged and there is no new discussion of SLS processes, once your final 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.

[3] For overseas researchers you will be required to skype with your SO at a time convenient for them.

Any changes in tables, figures, text or content requires another Final Outputs Clearance Form. For example, if you need to submit to a different journal then you should submit a Final Outputs Clearance Form for each submission. ***For speed of clearance, you should use 'track changes' if your output is in 'Word' format or 'compare and combine' for outputs in 'Powerpoint' (see image below) to ensure the SO can quickly see the changes.***



If you do not do this, clearance will inevitably be delayed as the SO will need to check for changes manually. As you have gained either SLS Approved Researcher status or Provisionally Approved Researcher status, the SLS trust that you will inform us of any changes and not release output which has not been cleared especially that which may inadvertently cause reputational damage to the SLS/NRS/Scottish Government. Releasing uncleared data would be a breach of the terms and conditions of your Undertaking with the SLS.

You **MUST** notify the SLS Support Team when any publication which draws on the SLS is published. We maintain a library of all published research that uses the SLS and it is vital for us to keep this up-to-date.

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.

Final Outputs Clearance Criteria

Describing the SLS (see LSCS Working Paper 1.0 for further details)[4]:

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 SO linked your datasets using the postcode (deleting the postcode before providing you with the data) then you must state that SLS staff carried out the linkage and not say that 'we' linked the data.

[4] 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 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.

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) it may be permissible to report on N-values between 5 and 10, however this will be at the discretion of your SLS SO. Thus, the minimum N of 10 also applies to tables and summary statistics.

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 SO 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 SO 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 Intermediate and Final 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 SO believes that data may be disclosive they must bring this to the attention of the SLS data custodian who will decide on the procedure to follow. In most cases this will require further aggregation of the data.

Annex to the SLS Undertaking Form

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 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:

1. For a first offence, the penalty should be a minimum 12 month discretionary suspension from access to any SLS or NRSS data applicable to the individual(s) in question. It would generate a written warning to that individual’s institution of employment.

2. An individual’s second 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).

3. 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.

Annex B Intermediate Outputs Clearance Criteria

Project Number:_________________ Researcher Name: ________________________________

File Name: ___________________________________

Where is your file stored?:

__Reports\To_be_cleared\____

(eg \Reports\To_be_cleared\2018_0XX_(1) _dd/mm/yy).

Intermediate Outputs Clearance Criteria CHECKLIST

As far as possible, limit intermediate output requests to one of the following scenarios:

1) Pre-publication i.e. necessary to produce your final 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.

Checklist for intermediate output

□ 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.

□ Tables, charts, variables and variable codes are 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

□ The sample used in each table/chart/model is shown i.e. description including exclusions and sample size (N)

□ Any graph/chart/percentage should also include the data behind the graph/chart/percentage, have no outliers and be in a suitable format. Scatter plots are not released.

□ No tables have a cell count of less than 10

□ Any statistics should be based on a sample size of at least 10

□ There is no disclosure between tables (differencing) in this output or previous outputs

□ Save this completed Intermediate 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, the SO will cease output checking and inform the researcher that they will need to return to the safe setting to amend the output

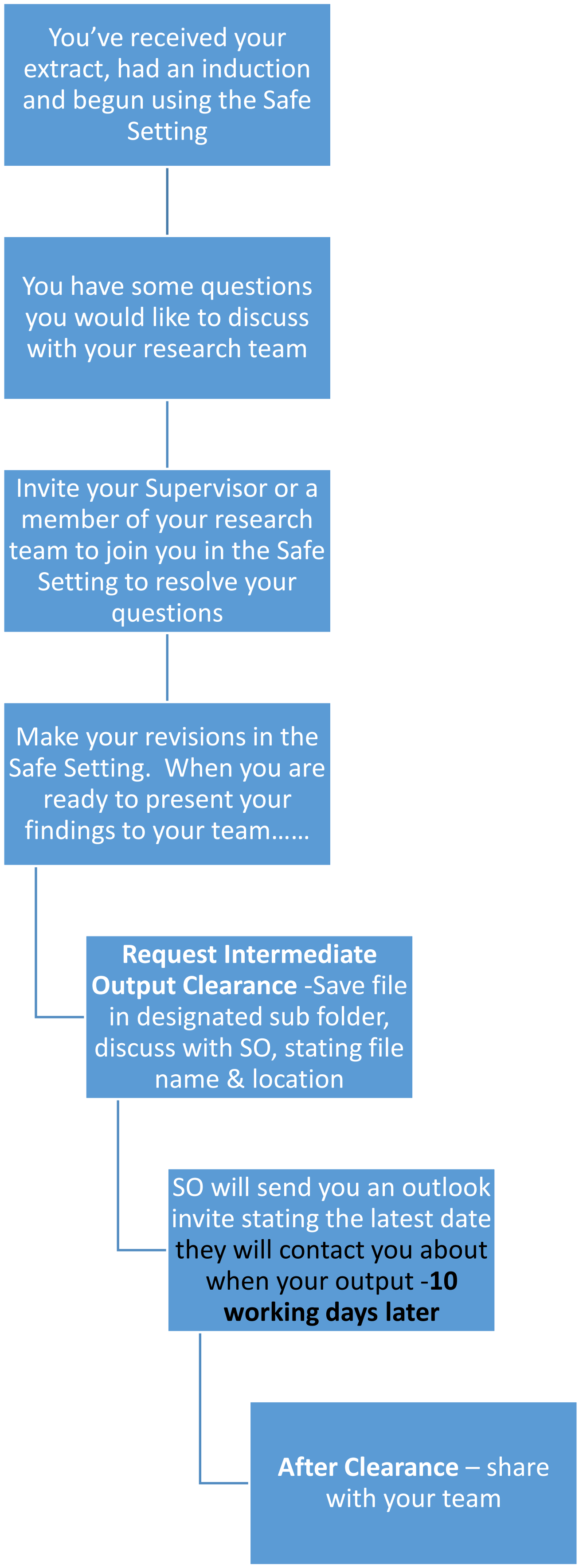
Annex C Final Outputs Clearance Criteria

CHECKLIST Final 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 SO) □
- 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 Intermediate output clearance



You've received your extract, had an induction and begun using the Safe Setting

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

Invite your Supervisor or a member of your research team to join you in the Safe Setting to resolve your questions

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

Request Intermediate Output Clearance -Save file in designated sub folder, discuss with SO, stating file name & location

SO will send you an outlook invite stating the latest date they will contact you about when your output -**10 working days later**

After Clearance – share with your team