A guide for researchers requesting outputs from the National Safe Haven

Document Control

Document Control	
Version	A guide for researchers requesting outputs from the National Safe Haven
Date Issued	12/10/2018
Author(s)	eDRIS Team
Other Related Documents	Public Health Scotland (PHS) Disclosure Control Policy
Comments to	phs.edris@phs.scot

Document History			
Version	Date	Comment	Author(s)
1.0	12/10/2018	First Publication of Disclosure Guidance	Dionysis Vragkos Carole Morris
1.1	15/10/2018	Minor revisions following feedback from research community	Carole Morris
1.2	23/11/2020	Updated reference to Disclosure Control Policy to reflect move to PHS; updated email address to PHS email; now includes reference to additional data controller SDC checks and authorisation; link to Secure Data Access Professionals SDC guide	Suhail Iqbal
1.3	05/01/2021	Updated acknowledgment replacing NSS to PHS	Diane Rennie

A guide for researchers requesting outputs from the National Safe Haven

Contents

A guide for researchers requesting outputs from the National Safe Haven

Overview

On behalf of the Data Controllers, eDRIS has a responsibility to manage 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 outputs requested for release from the National Safe Haven (NSH).

This document provides guiding principles that should be followed prior to requesting the release of such files.

Outputs are assessed based on data controller's disclosure control protocols. Should no other protocol be specified, outputs are processed with respect to *PHS' Statistical Disclosure Protocol*. https://www.publichealthscotland.scot/media/2707/public-health-scotland-statistical-disclosure-control-protocol.pdf

N.B. If your project includes data where the data controller has not delegated authority to eDRIS to carry out checks, for example NRS Census data, then statistical disclosure control checks and authorisation will also be carried out by the data controller.

General Guidance – Summary

This section provides a short summary of key points for researchers to consider prior to requesting a file to be released from the NSH. Further details according to the types of outputs can be found in the sections that follow.

- Processing time
  All outputs are checked by two members of eDRIS staff before being released. Please allow us sufficient time to review and assess your output. We aim to turnaround a request within 3 working days. However, this could vary dependent on the quantity, staff availability and complexity of tables, charts or other files in your request. If your output requires additional checks by data controllers, such as NRS for Census data, then please allow for more time to process your request.

- Clear outputs
  Please ensure that all your outputs are comprehensible enough to be reviewed by someone less familiar with the project. The NSH provides most of the tools required to produce outputs in their final format for a paper or publication. Such outputs are clearer and therefore less likely to require changes or further explanation.

- Small numbers
  Please consider whether counts of less than 5 are included in any of the requested outputs as these could lead to the re-identification of individuals. In some cases, (e.g. the output contains sensitive information) no cell counts of less than 10 should be included.

A guide for researchers requesting outputs from the National Safe Haven

- Differencing from previously released outputs
  As a project progresses, more and more outputs will be created. You should consider whether multiple tables may provide the ability to potentially identify individuals from differencing of 2 or more tables.

- Multiple individual requests vs. 1 request for multiple outputs
  If you have or expect to have multiple outputs requested for release within a short period of time, consider grouping them together (within reason) in one disclosure request so that we may undertake an assessment in respect of all the above points.

- Draft vs. Final output
  Outputs can be in the format they are going to be shared (e.g. in a publication) or in a draft format for further processing outwith the NSH. In case of the latter, you should not share the released outputs outwith the research team and in some cases you might be asked to confirm that in writing.

- Do you need it released?
  Do want a draft output to be released so that you can discuss with your project team? You could consider reviewing it within the NSH alongside any members of you project team that are named in your governance documentation and have completed relevant IG training.

Following assessment of your outputs, you may be advised that:

- Output may require some adjustment to minimise disclosure risk; for example, amalgamate categories to remove small numbers/differencing. Once you have done this, your output will require a further check – please allow sufficient time for this.
- Output may be released as final and can be published in the public domain.
- Output may be released for internal use by the project team only.

A guide for researchers requesting outputs from the National Safe Haven

Formats of Output

Examples of outputs include

- Descriptive statistics presented in table or chart formats
- Document formats (reports or papers)
- Statistical model outputs
- Coding files (syntax)

The released outputs can be either final format i.e. as they are going to be sent for review before publication or presentation in the public domain or in a draft/interim format that will be further processed outwith the NSH or used for discussion with the project team. Output requested should be in a clear a format as possible. Aggregated Data tables in a draft/interim format may be released for internal use of the research team only – you should indicate on your request if output is requested for this purpose. This can be retained but cannot be published in the public domain. Any variation of these interim outputs you do wish to publish should be checked by eDRIS prior to publication. Raw data, subsections of data, raw statistical output (e.g. raw STATA logs) and any data at individual level will not be released.

Requirements for release of outputs from NSH

All types of output

When raising a request for a file to be released from NSH, the researcher must always make sure that:

- Each file has a clear title and description.
  - File names should include the project number, date and version – a sensible naming convention can help the RC to process multiple outputs in one request.
  - Your request for release should include a short but thorough description of what each file presents – population size, range of data, demographic characteristics/parameters considered etc.
  - Your request should also detail what the output will be used for – publication, conference, further analysis, discussion with colleagues etc.
- The outputs are in a readable format and do not identify any individuals. Preferred formats are:
  - Charts of summary data (that do not display any disclosive information) that are clearly labelled
  - Formatted, labelled tables
  - There should be no embedded links or data
  - If using R, R Markdown script is a useful tool.

A guide for researchers requesting outputs from the National Safe Haven

Tables / Charts

If tables /charts are required, ensure that for any table within the output:

- You have considered the disclosiveness of small numbers (<5 or in some cases <10)
- There are no columns or rows dominated by zeros.
- No hidden columns or rows
- Minima and maxima are avoided as these represent the extremes and could identify individual cases. Where possible, present means, medians, inter-quartile ranges and standard deviations instead.

For any final output requested for release for the same study - to avoid identification through differencing between tables / charts:

- The same nested variable breakdowns or groupings for the same population and parameters should be used every time.
- The same period/ range of data should be used every time.

Example

If you have a file released with a summary per age group and your present findings for e.g. age group [16-25], you should not request similar information to be released in the future for age group [16-24] as there is a higher risk for people of age 25 to be identified – we recommend deciding how age groups will be combined and retaining those groupings for all the outputs.

In general, tables containing cell counts <5 will not be released. If your table has counts <5 consider if:

- The table can be redesigned (e.g. amalgamate categories).
- The <5 cells could be suppressed. If suppression is applied, you need to ensure that the suppressed cells cannot be derived from information given within any table in any output (e.g. subtracting existing values from the totals or differencing between tables). i.e. you may need to suppress more than one cell in a row or column.
- Table redesign is preferred over cell suppression.

If the output contains sensitive information, or geography is at a low level, counts fewer than 10 may require suppression or redesign. Counts from 5 – 9 will not be released if they:

- Refer to a small geography – e.g. individual island Boards, below NHS Board level, local authority, Hospital level or below
- Associate with 1 or 2 medical practices, practitioners, schools etc.

Remember that most of the time sensitive information can be presented in a different way and still deliver the required results without having to provide exact counts or tables.

A guide for researchers requesting outputs from the National Safe Haven

The examples below demonstrate different ways to show the breakdown of 3 different groups of people (population 1, 2, 3) with specific characteristics per gender. The table 1 includes numbers less than 5 and would generally not be released. Turning this into percentages or a graph could depict the same information with a reduced risk of identification.

Note that providing a total of rows/ columns in any of the tables below could still lead to identification and thus should be removed.

Table1 – *Potentially not good*: Counts for 3 different subgroups by age (jn years) and gender

Population Count by Gender

	Population 1	Population 1	Population 2	Population 2	Population 3	Population 3
	Male	Female	Male	Female	Male	Female
0-15	0	1	1	2	0	0
16-20	27	49	32	68	48	141
21-25	118	157	138	68	220	141
26-30	200	192	168	118	123	191
31-35	240	192	168	135	265	224
36-40	270	237	227	173	254	149
41-45	275	249	256	165	201	266
46-50	265	267	284	266	191	241
51-55	225	256	262	288	249	224
56-60	179	200	274	308	105	149
61-65	133	145	150	255	163	145
66+	4	6	21	52	12	33
TOTAL	1936	1951	1981	1898	1831	1904

Note: The cells where there may be disclosure concerns are highlighted in yellow.

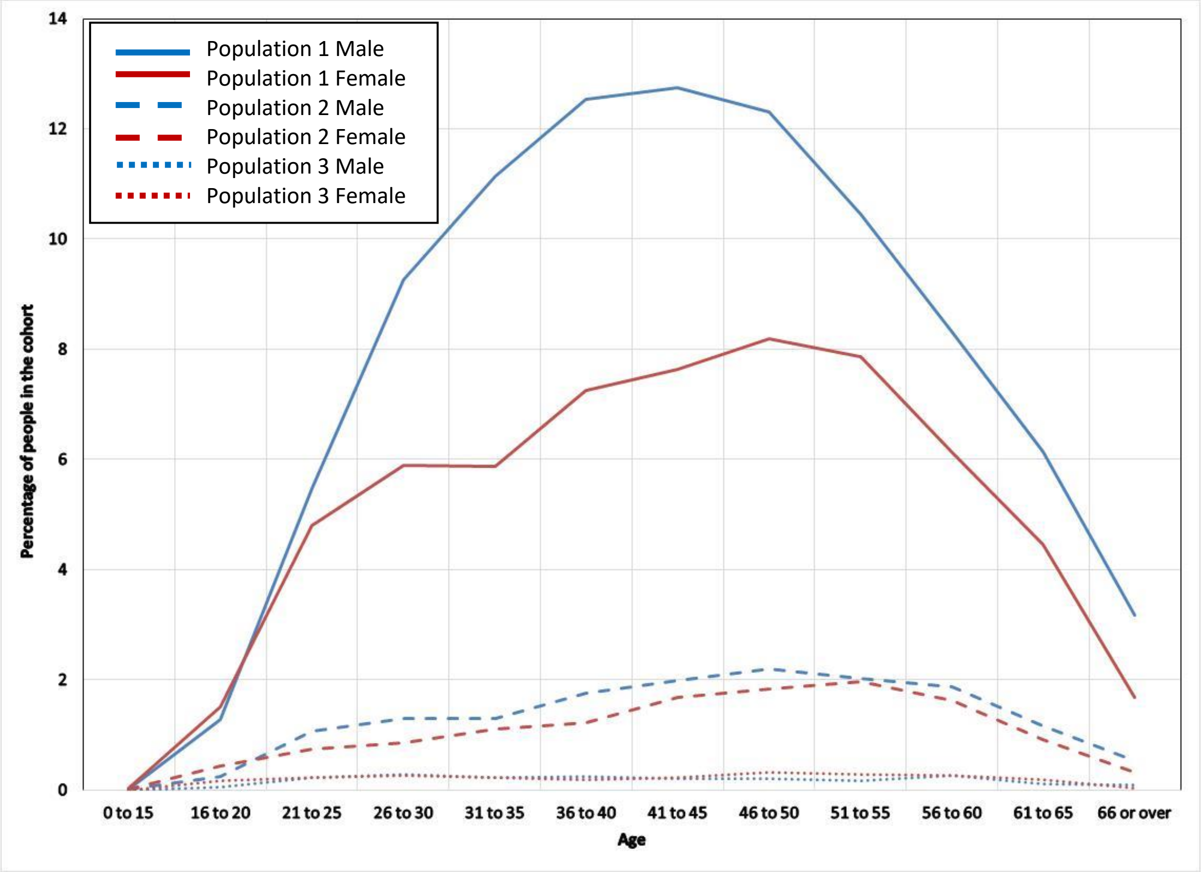
Table2 - *Better*: Percentages within gender for 3 different subgroups broken down by age

	Population 1	Population 1	Population 2	Population 2	Population 3	Population 3
	Male	Female	Male	Female	Male	Female
0-15	0.0	0.1	0.1	0.1	0.0	0.0
16-20	1.4	2.5	1.6	3.6	2.6	7.4

21-25	6.1	8.0	7.0	3.6	12.0	7.4
26-30	10.3	9.8	8.5	6.2	6.7	10.0
31-35	12.4	9.8	8.5	7.1	14.5	11.8
36-40	13.9	12.1	11.5	9.1	13.9	7.8
41-45	14.2	12.8	12.9	8.7	11.0	14.0
46-50	13.7	13.7	14.3	14.0	10.4	12.7
51-55	11.6	13.1	13.2	15.2	13.6	11.8
56-60	9.2	10.3	13.8	16.2	5.7	7.8
61-65	6.9	7.4	7.6	13.4	8.9	7.6
66+	0.2	0.3	1.1	2.7	0.7	1.7
TOTAL	1936	1951	1981	1898	1831	1904

Where the cohort/population is known the percentages will not mask the small numbers as they can be calculated using other values in the table. In such cases secondary disclosure control may be required by removing a second value so the true value is not known for 2 categories. Alternatively, the output could be presented as a chart if this successfully disguises the true value. Where the cohort/population is known the percentages will not mask the small numbers as they can be calculated using other values in the table. In such cases secondary disclosure control may be required by removing a second value so the true value is not known for 2 categories. Alternatively, the output could be presented as a chart if this successfully disguises the true value; see Chart 1 below which uses a line chart to show the distribution of different groups of individuals in the cohort by age without referring to exact numbers or percentages – this adds a layer of fuzziness without obscuring the overall pattern of the distribution.

Chart1 - *Best*: Percentage of people in the cohort by population type by age



You should always make sure that any graphical output does not link to the source data. For example, when copying an excel graph into a word document make sure that the actual worksheet is not embedded in the graph.

Statistical models

It is appreciated that sometimes statistical output files are required for interim outputs. These should always be clearly labelled or annotated to assist in the review and ideally they should be formatted into a document (e.g. Word).

Output for statistical models will be assessed using similar rules to the above. In summary, each statistical model should:

- Have clear titles and descriptions (including population it relates to)

- Sufficient observations for the model or at least 10 residual degrees of freedom as appropriate
- No residuals or residual plots that show individual pointsCoding files (syntax)

Output for syntax files will be assessed when:

- The code is clearly annotated with comments to assist the review
- No references or figures are included in the comments or code that could lead to potential identification of individuals
- No pseudo anonymised ID numbers are included in the code or the comments
- Minimise the volume of code to that which you actually would require to ease the review

Outputs for project team internal use only

Any outputs in a draft format, for discussion within the research team (named on your approvals and with relevant IG training), provided that they comply with the rules above, may be released as management information only. You should indicate on your request if outputs are requested for this purpose.

In very limited circumstances, cell counts of less than 5 could be disclosed for management information only if:

- They include non-sensitive information
- They refer to a large geography (no Island Boards or below, no single practices or below, etc.)

In such a rare occasion, the researcher requesting the output should **confirm in writing to their RC that circulation is restricted, that the output will be destroyed after x date and that individual confidentiality is not compromised.**

In addition, the researcher should acknowledge the following:

“This information has been released for management information purposes only. The data have not been adjusted to protect against potential disclosure risks and may contain information which enables (perhaps with the aid of further knowledge of the topic) an individual patient or member of staff to be identified.”

A guide for researchers requesting outputs from the National Safe Haven

Acknowledgements in publications

The eDRIS team ask that you acknowledge the use of our service and the National Safe Haven in any publications/presentations where appropriate. An example of this is shown below:

“The authors would like to acknowledge the support of the eDRIS Team (Public Health Scotland) for their involvement in obtaining approvals, provisioning and linking data and the use of the secure analytical platform within the National Safe Haven.”

Further Reading

Other Guidance Documents:

Scotland Census:
https://www.scotlandscensus.gov.uk/disclosure-control

Government Statistical Service:
https://gss.civilservice.gov.uk/guidances/methodology/statistical-disclosure-control/

Office for National Statistics:
https://www.ons.gov.uk/methodology/methodologytopicsandstatisticalconcepts/disclosurecontrol

Safe Data Access Professionals SDC guide:

https://securedatagroup.org/sdc-handbook/

Books:

Hundepool, Anco et al: Statistical Disclosure Control, John Wiley & Sons, Ltd, 5 July 2012. Print ISBN:9781119978152

Duncan, George T et al: Statistical Confidentiality, Principles and Practice, Springer, 2011. Print ISBN 978-1-4419-7802-8