|  |
| --- |
| http://www.abdn.ac.uk/iahs/slideshow-images/DaSH_logolong2_rdax_356x267.jpg**Grampian Data Safe Haven (DaSH)** |
| **Document Type /No.****DaSH10\_20** | **Version: 03** | **Information Classification:****Sensitive** |
| **Output releases and disclosure checking: Guide for DaSH researchers** | Author: **Vicky Munro***Person Responsible for creating the document* | Approved By:**Antonietta Chaliou***Co-signatory* |

**This Document will be reviewed at least every 3 years from the approval date.**

**Document Revision History:**

|  |  |  |
| --- | --- | --- |
| **Version** | **Approval Date** | **Description of Change** |
| 01 | Refer to Q-Pulse | Document creation |
| 02 | Refer to Q-Pulse | Minor edits to ‘Minimising Disclosure risks’ |
| 03 | Refer to Q-Pulse | Updated Authors |

All printed copies are **UNCONTROLLED**. Please ensure that you are working from the current version of this document.

# Output releases and disclosure checking: Guide for DaSH researchers

## Why is disclosure checking of safe haven outputs important?

Safe Havens are secure environments supported by trained staff and agreed processes whereby health data can be processed and linked with other health data (and/or non-health related data) and made available to approved researchers in a de-identified form to facilitate research. Safe havens provide an environment whereby robust control and safeguards mitigate the risk of compromising patient privacy and confidentiality associated with such research.

Researcher responsibilities for conducting research within the Grampian Data Safe Haven (DaSH) are outlined in the generic Data Management Plan (gDMP) and Investigator Declaration**. It is the responsibility of the project Principal Investigator (PI) to ensure all team members are familiar with the gDMP and their responsibilities when accessing data within DaSH.** This is crucial to avoid data breaches.

DaSH defines a breach as **“a breach of security leading to the accidental or unlawful destruction, loss, alteration, unauthorised disclosure of, or access to, personal data transmitted, stored or otherwise processed.”** DaSH Users are responsible for reporting any inadvertent events that they have caused or are aware of that are in breach of the conditions in the generic and / or specific Data Management Plan. Users should report to DaSH via email (DaSH@abdn.ac.uk) in the first instance.

Any breaches will be recorded, investigated, and dealt with according to the DaSH breach procedure. Further upward escalation of the breach will be as per the standard procedures for NHS Grampian and University of Aberdeen. **Breaches are taken very seriously and, following investigation, can potentially lead to disciplinary action if considered negligent or malicious.**

## Process for releasing data from the safe haven

Researchers who wish to release the results of analyses or any other project outputs (e.g. code) from their DaSH folder should send the following information to DaSH@abdn.ac.uk and one of the DaSH Research Coordinators (RCs) will carry out a review:

* DaSH project number
* File names and location (including sub folder names if relevant) – *ideally all files will be in the same sub folder (e.g. DaSHXXX > For extraction)*

We’d also encourage you to provide the following information in your email request to help facilitate the review process:

* Description of the files *(e.g. summary report in Word for journal submission; R code for future use; descriptive statistics in Excel for research meeting, etc.)*
* Population / cohort information
* Any disclosure concerns (e.g. potential small numbers) and justification – *please remember that data should not be copied out of the safe haven so should not be included in the email request*
* Other relevant information *(e.g. if requesting the release of hypothetical cases or predictions from modelling, this should be clearly stated and explained in the request email and output file so that the RCs can distinguish between true patient level data / small numbers and predicted data)*

**TIP:** A researcher Output Request Checklist can be found in Appendix A.

Although DaSH generally releases outputs within two working days, please allow longer if you require the outputs for specific deadlines as timescales will depend on the nature of the files requested and staff availability. It is also possible that any disclosure concerns will delay release until these are resolved.

**Please note that a request to DaSH is the only way that results/outputs and information should be removed from the safe haven folder (i.e. researchers must not transfer or copy data by any means, including via screenshots, printing, email, or written notes). Other methods of removing data (e.g., including writing data or pseudonymised IDs in a notepad or referring to in an email) would be in breach of the Investigator Declaration.**

## General release criteria

Researchers should check their Data Management Plan (DMP) and approvals for any project-specific release criteria, however the general rules are:

1. No individual patient level data should be included in files for release (i.e. no patient data copied directly from your research datasets)
2. Files should not include any cells containing less than 5 individuals

## File preparation

Prior to requesting a release, researchers should ensure that files are in a format suitable for a report or publication if possible. This includes:

* **Aggregate data:** Ensure that no patient level data is contained within the files as only aggregate data can be released. This also applies to graphs (e.g. scatterplots in which each point represents one individual would generally not be released[[1]](#footnote-1)).
* **Labelling:** Ensure all tables / graphs are labelled appropriately so that DaSH RCs know what the variables are and what the data represents. The outputs should be understandable by those not familiar with the project.
* **Sample sizes:** Ensure that sample sizes are clearly stated so that DaSH RCs can identify any small numbers (i.e. < 5). For example, total counts should be provided if percentages are being released. The total number of individuals included in a statistical analysis model should also be provided for regression analyses, etc. to ensure these are not based on small numbers.
* **Minimise data for release:** Only data that is required should be released (i.e. large SPSS output files including data not needed for analysis / reporting purposes should not be requested for release). For initial analysis, you should consider whether the project team (i.e. those with the required approvals) can review the data within the safe haven environment rather than requesting for release.
* **Multiple requests:** If you anticipate multiple requests for releases within a short period, please consider whether these can be requested via one release rather than separate requests.
* **Previous outputs:** You should consider whether the data in your requested output will potentially be identifiable if combined with data available in previous releases. It is the researcher / PI’s responsibility to ensure that this is not the case.

## Minimising disclosure risks

Prior to requesting an output for release, you should undertake disclosure checks with the aim to identify and minimise disclosure risks (i.e. the risk of personally-identifiable information being released). DaSH RCs will also undertake disclosure checks prior to release and may request changes if any risks are identified.

**Please note that it is the Principal Investigator’s responsibility to ensure that patient confidentiality is not put at risk. Therefore, the files will be released under the condition that the appropriateness of presenting or publishing the outputs is considered.**

If an output contains small numbers (i.e. <5), researchers should consider adjusting their analysis using one of the following approaches:

* Remove low cells values (i.e. suppression)
* Replace it with an indicator (e.g. <5)
* Convert it from a count into a proportion *– please note this may not be appropriate if the number can still be calculated*
* Round the counts (e.g. to the nearest 5)
* Redesign tables (e.g. combine categories)

If low counts are removed or replaced with an indicator, you should ensure that they cannot be calculated using totals or other data present in the outputs (including previously requested outputs).

If small numbers (<5) are deemed necessary for analysis or reporting purposes, this would require sign off by the project PI, DaSH Leads and/or data custodians and permission providers.

## Advice on specific file type and formats

### Tables / Charts

Researchers should ensure there are:

* No patient level data or cells with small numbers (i.e. <5)
* No hidden columns, rows or sheets
* No ranges where the min or max values represent the extremes and could identify individual cases. Instead, consider using medians, means, inter-quartile ranges or standard deviations

### Html files

Researchers can request release of html files generated from R Markdown. For very small files (i.e. with fewer than 50 lines of source code), the html file and source code will be reviewed and released.

For larger files, the html file needs to be knitted in the virtual presence of a Research Coordinator (by arranging a MS Teams call and screen sharing) and checked alongside the source code before being released. The code used to generate html files needs to be clear and well annotated so that it can be reviewed by the Research Coordinator.

When emailing DaSH to request an html output for release, please include the file names and location of both the html file and the code (.rmd) used to generate the files. Please note that the code will be used for checking purposes only unless the researcher states that they also wish for this to be released. Once initial checks are completed, a Research Coordinator will contact the researcher to arrange a Teams call to observe the researcher knitting the file prior to final disclosure checks and release. **Please note the disclosure checking process may take longer for this file type.**

Before the meeting, researchers should copy or move the previously generated html file and ensure that if the location of the .rmd file has changed (to a dedicated ‘outputs’ folder) it can still be knitted (data sources need to be hard coded in such case).

### Code / Syntax

Researchers requesting the release of code / syntax should provide a written explanation of what the code / syntax is, as well as confirmation that this does not contain any identifiable information.

In addition, researchers should ensure:

* Identification of individuals from the code is not possible
* No pseudo-anonymised ID numbers are included
* Only code required outside the safe haven is requested for release

### Image files and graphs

Patient levels images (e.g. DICOM images) will not be released from the safe haven. Irrespective of format (e.g. PNG files, Stata graph files, PDFs, etc.), graphs will only be released if they meet the usual criteria (e.g. no small numbers or patient level data). Therefore, aggregate data is likely to be considered for release but scatterplots will generally not.

Researchers should provide sample sizes on graphs to support the disclosure checking process. If required, DaSH Research Coordinators may request the tables used to create the graphs in order to complete the disclosure checks.

## AI Models

Processes for AI model disclosure checking are currently under development. If these are requested for release, DaSH will initially ask for confirmation that their research outputs:

1. Contain no generative AI models or other means of creating or recreating NHS patient or confidential data.
2. Have been thoroughly examined by your organisation using an auditable quality management process, to remove all patient and confidential data prior to export (including, but not limited to executable code, non-executable code, code libraries and frameworks, code comments or notes, databases, reference files, media files, executable/binary files, Jupyter notebooks or an equivalent, models, etc.).
3. Contain no patient or confidential data (synthetic or real, deidentified or otherwise, encrypted, obfuscated or otherwise) or small counts (i.e., frequencies < 5)
4. To the best of your knowledge, cannot be used to reverse-engineer deidentified, encrypted or obfuscated patient or confidential information
5. Are compliant with the terms and conditions of any access to data

A call will then be arranged with the researchers to talk through their model and what this contains. The model will then require sign off for release by the DaSH Technical and/or Clinical Lead. At this point researchers will be asked to produce a check digit within the environment and this will be provided at the release stage.

## Disclosure control examples

We’ve selected some examples from the **Handbook on Statistical Disclosure Control for Outputs**, produced by The Health Foundation and partners, to highlight potential disclosure issues and give insight into the type of recommendations featured in the handbook. For further information and additional examples, please refer to the handbook.

The handbook uses the following traffic-light style coding to illustrate what data is / is not likely to be released:


### Descriptive statistics

###



**Tips for researchers wishing to release descriptive statistics:**

* Provide understandable labels to assist with the disclosure checking
* Consider grouping categories to avoid small cell frequencies (i.e. < 5)
* Avoid including minimum and maximum values which could be based on one observation

### Kaplan-Meier curve



**Tips for researchers wishing to release survival plots:**

* Provide overall sample size for figure in release file
* Refer us to the accompanying table so DaSH RCs can review the underlying data to check for small frequencies. This should be a separate file for checking purposes only rather than part of the output file. Depending on the underlying data, this may require sign off by PI, DaSH Leads and/or data custodians

### Regression



**Tips for researchers wishing to release regression results:**

* Ensure that the regression is not undertaken on a single unit (or <5) – you should include the number of observations in your output so this can be checked
* If undertaking sequential regressions, the number of observations in the cohort should not change by only a small number each time
* Ensure that the regression does not consist solely of categorical variables (i.e., 0/1, True/False, Yes/No, etc.)

### Margin plots



**Tips for researchers wishing to release margin plot graphs:**

* Outputs must include the degrees of freedom and number of units used to produce the model for DaSH Research Coordinators to review (i.e. these should be at least 5)
* Models should not be based on one unit

## Acknowledgement in publications

DaSH request that you acknowledge the use of the safe haven in any publications or presentations. For presentations we can provide you with a slide to include. For publications, example text is below:

*The authors acknowledge the support of the Grampian Data Safe Haven (DaSH) facility within the Aberdeen Centre for Health Data Science (ACHDS), and the associated financial support of the University of Aberdeen and NHS Research Scotland (through NHS Grampian investment in DaSH).*

## Further reading

DaSH recommend the following reading:

*Handbook on Statistical Disclosure Control for Outputs (2019):* [*https://securedatagroup.org/sdc-handbook/*](https://securedatagroup.org/sdc-handbook/)

## Further support

If you have any questions or would like further advice on disclosure checking and preparing files for release, please contact DaSH@abdn.ac.uk

# Appendix A: Output request checklist

|  |  |
| --- | --- |
| **Output request stage** | **Complete** |
| **Prepare file for release (see page 2-4):**Ensure that any files required for release are as close to publication or report format as possible. This includes:* **Aggregate data:** Ensure that no patient level data is contained within the files as only aggregate data can be released. This also applies to graphs (e.g., scatterplots in which each point represents one individual would not be released).
* **Labelling:** Ensure all tables / graphs are labelled appropriately so that DaSH RCs know what the variables are and what the data represents. The outputs should be understandable by those not familiar with the project.
* **Sample sizes:** Ensure that sample sizes are clearly stated so that DaSH RCs can identify any small numbers (i.e., < 5). For example, total counts should be provided if percentages are being released. Counts should also be provided for regression analyses, etc. to ensure these are not based on small numbers.
* **Minimise data for release:** Only data that is required should be released (i.e. large SPSS output files including data not needed for analysis / reporting purposes should not be requested for release). For initial analysis, you should consider whether the project team (i.e., those with the required approvals) can review the data within the safe haven environment rather than requesting for release.
* **Multiple requests:** If you anticipate multiple requests for releases within a short period, please consider whether these can be requested via one release rather than separate requests.
* **Previous outputs:** You should consider whether the data in your requested output will potentially be identifiable if combined with data available in previous releases. It is the researcher / PI’s responsibility to ensure that this is not the case.
 |  |
| **Consider and minimise disclosure risks:*** Review output to ensure it meets the criteria for release (i.e., no small numbers or patient level data; see page 3-4)
* Consider any additional risks associated with output or analysis type (see pages 4-10)
* Minimise any potential disclosure risks (see page 4-10)
* Discuss with your PI for advice if appropriate
 |  |
| **Request release (see page 2-3):**Email DaSH@abdn.ac.uk with the following information:* DaSH project number
* File names and location (including sub folder names if relevant) – ideally all files will be in the same sub folder
* Description of the files (e.g., summary report in Word; R code; descriptive statistics in Excel, etc.)
* Population / cohort information
* Any disclosure concerns (e.g., potential small numbers) and justification – *please remember that data should not be copied out of the safe haven so should not be included in the email request*
 |  |

1. Scatterplots from statistical models plotting predictions versus model residuals may be reviewed on a case-by-case basis. [↑](#footnote-ref-1)