SWANSEA UNIVERSITY MEDICAL SCHOOL

# OUTPUT REVIEW POLICY

| | |
|---|---|
| Version | 1.7 |
| Issued By | ASHLEY AKBARI |
| Classification | OPEN |
| Review Date | 06/06/2025 |
| Status | APPROVED |
| Reference | SAIL-POL-024 |
| Note | UNCONTROLLED WHEN PRINTED |

## CONTENTS

1 Overview
2 Standards for Output Control
3 Standards for Researchers
4 Standards for Reviewers
5 Appeals and Service Promise
6 Compliance

Appendix 1 – Simple Request Definition
Appendix 2 – References and further reading
Appendix 3 – Documentation Management

## PURPOSE

This document will detail the SAIL Output Review Policy. This policy is to be used in association with the SAIL Output User Guide, SAIL Output Review Checklist and the SAIL procedure for Output Review Training.

## SCOPE

This policy applies to all output requests made to SAIL. No SAIL staff/resource is exempt from this policy.

## 1. OVERVIEW

Review of outputs is a key part of the controls that the Secure Anonymised Information Linkage (SAIL) Databank uses to ensure the safe and legal use of anonymised data for research. All outputs produced based on data held by SAIL are manually inspected to ensure their appropriateness for release.

Output review has two main goals:

- Prevent the release of potentially disclosive results—meaning results with the potential to reveal information about a known individual.
- Ensure that the data is used only in line with SAIL policies, agreements with data providers, and the specific approval that is in place for the use of the data.

The goal in implementing this process is to maximise the utility of the data for research, while safeguarding the privacy of the individuals whose information we hold, as well as safeguarding our ongoing operations. The risks we are seeking to manage are:

**Breaching individuals' privacy –** information about a known individual is revealed (spontaneous recognition – see Disclosure by spontaneous recognition).

**Legal violations –** SAIL's legal basis for operation is a judgment that our safeguards and processes ensure the anonymity of the data. If we are found to release potentially disclosive results (even if no individual is actually identified), we may be in breach of the law.

**Reputational** – SAIL relies on the goodwill of data providers, funders, and other stakeholders to operate. If our processes are perceived to be unsafe, then we risk losing funding or not receiving data and being unable to operate.

Personal data is defined under Article 4 of the General Data Protection Regulation (GDPR) as any information relating to an identified or an identifiable natural person. The SAIL Databank works to the standards set out

in the UK Information Commissioners Office 'Anonymisation Managing Data Protection Risk Code of Practice' to ensure that the information it holds is adequately de-identified during its lifecycle. Data that is adequately anonymised, e.g. through the application of this policy's measures, is not personal data and falls outside the remit of the GDPR and UK DPA 2018.

In accordance with the above, SAIL works to ensure that *all* outputs from SAIL are fully anonymised. SAIL procedures therefore, err on the side of caution in output review. To ensure legal compliance, this policy shall be applied consistently with the aim of maintaining and upholding anonymisation and minimising any risks to both the individual and the organisation.

SAIL operates a principles-based output review service (Ritchie & Elliot, 2015). Rather than having hard rules, we use a set of rules of thumb and rely on trained and experienced reviewers applying these rules of thumb and their judgment to ensure that outputs are safe (see S3 - Standards for Researchers).

This policy applies to standard SAIL projects where anonymisation is the legal basis enabling use of the data. Where there is another legal basis for data sharing and an explicit agreement in place with data providers, these restrictions on outputs may not apply (for example, for studies with consent to link data, some data controllers have agreed that SAIL can release individual patient data).

## 2. STANDARDS FOR OUTPUT CONTROL

In addition to the specific principles described below, reviews will be conducted with reference to the *Handbook on Statistical Disclosure Control for Outputs*, version 1.0 (Griffiths et al., 2019). The rules of thumb described in this handbook will serve as the starting point for any type of output not covered by specific guidance in this document.

**Rules of Thumb - The analysis is permitted by SAIL policy and relevant data sharing agreements and project approvals.**

The output review process exists not just to prevent disclosure of information about individuals, but also to ensure that SAIL data is used only in line with our policies and agreements with data providers specific instructions. A check shall be undertaken to establish whether specific instructions are in place from the relevant data provider/controller. Only research that has been reviewed and approved by the independent Information Governance Review Panel (IGRP) may be conducted. Other limited outputs are allowed by policy (such as feasibility checks and sample sizes, as described in the SAIL Access Policy).

Where outputs are requested under an IGRP-approved project, person-based data used to produce the results must be limited to what is approved to the project (including specified data sources and datasets, date and geographic coverages within data, and specified restricted data items, etc.). The analyses must fall within those described in the application, and any other specific restrictions to the output described in the IGRP application must be adhered to.

**Data should be processed inside the trusted research environment, with only results requested out.**

The expected operating model of SAIL is that all data preparation and analysis will be conducted within the SAIL Trusted Research Environment (TRE), with only results requested and approved out.

Outputs that consist of data in a form other than results suitable for review or publication (such as intermediate processed data for further processing outside of SAIL) will not be allowed, even if they otherwise meet the requirement for a safe output. Exceptions to this must be reviewed and approved by the IGRP.

Outputs are not restricted to final results. Intermediate results (for example, for review by the project team) are allowable, as long as they meet the requirements for safe outputs. However, researchers are strongly advised to conduct review of intermediate results within the SAIL TRE as it minimises risk, is more convenient

for researchers, and reduces reviewer workload. This can only be done if all researchers have SAIL accounts for the projects. Therefore, it is highly recommended that Projects should apply for sufficient Limited Access User (Read Only) accounts to allow for project review and inspection of outputs within the TRE.

**Individual-level data is not allowed out.**

No individual-level person-based results are allowed out of the SAIL Databank TRE for anonymised research studies. This is true even if the results can be shown to be non-disclosive.

**Cells shall not contain a value of less than 10.**

In accordance with the UK Information Commissioners (ICO) Guidance on Anonymisation it is acknowledged that small counts increase the risk of spontaneous recognition and the ability to cross-reference data when trying to reconstruct an individual's data. Normally, tabular and other results showing counts of fewer than 10 individuals, or events related to individuals, will not be allowed in accordance with the ICO guidance. Researchers may choose to mask these values (replacing the exact value with "<10" or "*", for example) or deal with the situation in other ways, such as further aggregating categories. It is the responsibility of the researcher to find solutions that adhere to this policy, not the responsibility of the output reviewers.

Cases where a count or difference of <10 may be derived are also disallowed. For example, if a table reports only one value less than ten, as well as a total, the exact value can be derived with trivial effort. Small numbers may additionally be derived from combining multiple tables, rates or percentages, etc., and such results are disallowed as well. These cases may be mitigated via masking the next-lowest count, further aggregation, or other changes to the results being reported.

Where many related tables are requested out, it may be difficult or practically impossible for a reviewer to assess whether any small numbers can be derived due to the large number of combinations that would need to be considered. Outputs may be rejected for this reason.

There are many examples of outputs that may contain small numbers, but arguably present little or no disclosure risk. However, exceptions to the <10 restriction are allowed only when a researcher can demonstrate that the outputs are both safe and necessary/important for release. Most of the time, the exact number in these cases does not provide statistically meaningful information and reporting that there were <10 cases is just as useful as reporting that there were 3 cases (for example). Thus, in the spirit of minimising outputs, as well as limiting risk and reviewer workload, exceptions are limited to those cases where the researcher can provide a compelling reason for release.

Reporting errors or missing data in cases where it does not plausibly relate to a recognisable characteristic of a person is not subject to the <10 rule.

**Totals**

Each output request must include all relevant totals and denominators to enable reviewers to perform consistency checks and verify small counts. Requests submitted without these totals may be rejected.

**Zero is disallowed where there is potential for disclosure.**

Reporting a count of zero is disallowed when, in combination with other reported information, it has the potential to disclose something meaningful about an individual or group of individuals. In most situations, this is not an issue; however, researchers should ensure that outputs, including zeroes, do not permit class disclosure.

**Maximum or minimum values are disallowed where there is the potential for disclosure related to outliers associated with a single individual.**

For many types of results, the minima and maxima are likely to be outliers that are related to a single individual. Where this is the case, and there is the potential of disclosing information about that individual, the reporting of minima and maxima will be disallowed, even if the exact number is not actually reported.

**Graphs and other visualisations are treated the same as numeric results, where exact values can be determined.**

Where an exact number can be derived from a visualisation, the small number rule and other rules will be applied as if it were a reported number. For example, if a bar chart reports a value <10, this would be disallowed.

Vector graphics files, such as .svg files, contain very high-precision information and may lead to small numbers being derivable, even if this is not apparent from viewing the graph.

Scatter plots, as well as box-and-whisker plots that include individual points for outliers, are typically disallowed because they display individual-level data. This includes plots of residuals.

Kaplan-Meier survival plots often feature drops in the line that relate to small numbers of individuals or a single individual. These small steps in the survival curve may be problematic, especially when the total group count or subgroup count is small. Therefore, in many cases, these small differences represent a disclosure risk. Censoring lines should always be excluded as they represent individuals. Any step in the curve must also represent 10 or more individuals. Reviewers will consider the disclosure potential of these plots on a case-by-case basis. To aid this review and risk assessment, all Kaplan-Meier survival plots must be accompanied by a risk/contingency table as a supporting document. Further advice on confidentialising survival plots is presented by O'Keefe et al. (2012)

**Nothing that could be used for performance tracking of individual organisations may be released.**

The data sharing agreements with data providers to SAIL prohibit the use of SAIL for the performance management of identifiable organisations, and this restriction is enforced through the output review process. Reporting of results that could be used for performance management of organisations is not allowed.

Outputs that report differences in outcomes between healthcare providers, data providing organisations, or geographic areas are allowed only when explicitly mentioned in a project's IGRP application, which has received approval.

**Machine Learning / AI Outputs**

Machine learning models present particular challenges for output checking. Researchers should engage with SAIL at an early stage of research to discuss the models to be used, likely outputs, and how those outputs will be utilised. These should be discussed at the scoping and IGRP application stages.

## Other Considerations

**Special requirements imposed by data providers.**

Occasionally, an individual data provider may have additional requirements or restrictions placed on the outputs from research using their data sources. Where this is the case, reviewers will apply these requirements **in addition** to the normal requirements outlined in this policy. For any result based on multiple data sources or datasets, including one with a special requirement from the data provider, this requirement will apply to the whole result. This applies to all outputs from projects provisioned with such datasets, regardless of which datasets are used in the outputs.

It is the responsibility of the project to adhere to any requirements data providers have regarding the sharing of outputs prior to publication.

SAIL will maintain a list of data sources and datasets with special review requirements for reviewers to use. This information will be made available to SAIL project applicants and users, ensuring they are aware of these requirements before applying for and using SAIL projects.

**File Format**

Hidden results or other hidden information (even if not results) is **not allowed**—for example, embedded Excel files within an Office document, macros, comments, formulas, or tracked changes.

Graphical information within documents from other sources/programmes must be shared as an image. This is to prevent links to underlying data from being left in documents, as this can provide access to hidden data. Pivot tables must be submitted as tables with links to underlying data broken.

Reviewers may request that outputs be resubmitted in another file format if hidden or embedded information (or the possibility of hidden or embedded information) makes it impossible or impractical to review the file.

**Code Files**

Code, scripts, and other files that don't contain data or results may be requested for publication or sharing outside of the secure environment, where necessary or beneficial. If the files are extensive, extra time may be required for review (timescales to be discussed on a case-by-case basis with reviewers).

Researchers should ensure that the files really have no outputs based on the data (for example, the result of a command mentioned in a comment within code); they should further take care not to include account passwords (it is not good practice for passwords to be saved in code files, in any case).

## 3. STANDARDS FOR RESEARCHERS

**Expectations**

It is the researcher's responsibility to produce safe outputs. Only files that the researcher has already checked and believes to be safe and in line with SAIL policies and project approvals should be submitted.

The responsibility of identifying the appropriate remedial steps to ensure safe output lies with the researcher, not with the Reviewer.

For resubmitted requests due to rejection, researchers must provide the rejected ID request and document how they have addressed the issues raised.

By submitting an output request, the researcher confirms that they have checked and believe their outputs to be safe. This means providing sufficient documentation for reviewers to understand the outputs, as well as answering questions or providing additional information where necessary to demonstrate that the results do not pose a risk of present or future disclosure (e.g., when combined with other information outside of the SAIL TRE) and do not contravene SAIL policies. Submission of outputs without adequate documentation will lead to immediate rejection.

All outputs reported from SAIL must be requested out via the output review process. Even small outputs, such as a single number that a researcher may easily remember and write down, should be requested via the output review system so that all outputs are documented and that both SAIL and the researcher have confidence that the output is safe.

Any deliberate attempt to circumvent the output review system constitutes a breach of the SAIL Data Access Agreement and will result in disciplinary action against the researcher.

**Documentation**

All output requests must be fully documented as a stand-alone request—not relying on information provided in previous requests or assumed knowledge of the reviewers.

Documentation should clearly indicate which data sources and datasets were used in the outputs. Each row and column in a table, as well as graphs, should be clearly labelled. The reviewers should have sufficient information to understand the methods, including the statistical tests used, as well as any relevant information regarding potential disclosure (such as whether the results are from a sample or a whole population, whether categories are exclusive, and whether a cell value could be derived from a total, etc.).

All requests should include confirmation that the results are in line with the <10 rule and other rules of thumb or should include justification that shows the output is safe and important to release if the researcher is requesting an exception to one of these rules.

**Minimising Requests**

Researchers should seek to request out only the results/outputs needed as few times as possible. Sometimes corrections or updates to analysis are inevitable. However, repeatedly requesting the same or similar results/outputs adds to the reviewer workload and increases exposure to the risk of disclosure (small numbers may be derived from combining multiple files).

If a set of outputs needs to be updated with an additional result, please request out only the new result. Appending the result to an existing document requires the entire document to be reviewed again.

**Conduct**

Researchers are expected to conduct themselves professionally during the Output Review process. The output reviews are to be treated like any other peer review.

## 4. STANDARDS FOR REVIEWERS

### Expectations

It is the responsibility of the Reviewer to: (1) only release results that they understand and (2) ensure that all aspects of each output adhere to all SAIL Policies.

Results should be reviewed in a timely manner; and decisions should not unduly restrict outputs that are safe and allowable, as this reduces the benefit derived from the data.

If the researcher has made a good faith effort to document the request, but the Reviewer does not understand it, it is the Reviewer's responsibility to seek help and/or clarification rather than reject the file.

In all cases where a file is rejected, a clear reason for the rejection should be provided. Normally, this would reference one of the principles outlined in this document.

### Dual Review

All output requests will be reviewed by two reviewers, with results released only where both reviewers are confident that it is safe. Each Reviewer is expected to conduct a full, independent review and have confidence in the entire request rather than relying on the other Reviewer.

If reviewers disagree, an additional senior reviewer will also review the request, and they will discuss the issue and attempt to reach an agreement. Outputs will be released only where all reviewers agree that it is safe. If one Reviewer still has concerns after this additional review, it will not be released.

## 5. APPEALS AND SERVICE PROMISE

### Appeals

If a researcher disagrees with a rejection and requests reconsideration, an additional senior reviewer will assess the request. When reviewers disagree (above), all three reviewers will discuss the issue, and the decision will be changed only if all three are satisfied that it is safe. The Directors will review all appeal decisions that result in a refusal.

### Service Promise

The time taken to undertake the initial review of outputs will be five working days after the date of submission. We will endeavour to complete simple requests (See Appendix 1 for the definition of simple requests) within five days, but this cannot be guaranteed. The result of this review will be a decision or a request for clarification if additional information is needed.

## 6. COMPLIANCE

Any staff subject to this policy who fail to comply with the provisions as set out above shall be subject to appropriate management response. This may include disciplinary action in accordance with the Swansea University Disciplinary Code and Procedures.

SAIL Information Security policies, standards, procedures and guidelines shall comply with legal, regulatory and statutory requirements.

## APPENDIX 1 – SIMPLE REQUEST DEFINITION

Simple request standardised file format requirements:

- For visualisations: common raster file formats (.png or .jpeg).
- For results: A single sheet MS Office Excel document (.xlsx) - 1 sheet with no relationship between values across rows or columns.

A SAIL output request can be considered a simple request if it abides by the SAIL Output Policy and only adheres to the following:

- All counts are rounded to the nearest multiple of 5 (with no counts <10), and all statistics are based on rounded counts.
- Percentages listed in their own cells.
- Descriptive Statistics: Means, modes, standard deviations, counts, sums, totals, or group-level summaries for large sample sizes with sample sizes clearly stated. Typically low disclosure risk unless working with small subgroups.
- Cross-tabulations: Simple two-way cross-tabulations with large sample sizes which are clearly stated.
- Plots: Graphical summaries which aggregate data visually and don't reveal data points relating to individuals. Histograms, Box Plots (with outliers removed and a confirmatory statement included), Line Plots, Venn Diagrams, Pie Charts, Area Plots, Principle Components Analysis Plot where shaded areas only are present
- Aggregated time series data: High-level trends over time (e.g., monthly averages) with sufficient aggregation to prevent small numbers from being derived.
- Statistical hypothesis tests: for large subgroups with clearly stated sample sizes.
- Basic correlation coefficients: for large sample sizes with these clearly stated.
- Basic multivariable regression outputs: Outputs from standard linear or logistic regression with large sample sizes, which are clearly stated. Odds/risk ratios, confidence intervals, and p-values are often straightforward as long as there are a large number of observations and the model is not over-fitted.
- Fitted curve parameters (standard deviation, skew, kurtosis).

## APPENDIX 2 – REFERENCES

| Document | Link |
|---|---|
| Green et al. (2024) *Setting up output checking processes - A guide for data services* | Guide to output checking processes |
| Griffiths et al. (2019) *Handbook on Statistical Disclosure Control for Outputs, version 1.0* | https://ukdataservice.ac.uk/app/uploads/thf_datareport_aw_web.pdf |
| Information Commissioner's Office (2012) *Anonymisation: managing data protection risk code of practice* | https://ico.org.uk/media/1061/anonymisation-code.pdf |
| O'Keefe et al. (2012) *Confidentialising Survival Analysis Output in a Remote Data Access System* | https://pdfs.semanticscholar.org/ac15/804b97aebc6bcea201ca23147a3e1f79c2d5.pdf |
| Ritchie & Elliot (2015) *Principles – versus rule-based output statistical disclosure control in remote access environments* | iqvol_39_2_ritchie.pdf |

# APPENDIX 3 - DOCUMENT MANAGEMENT.

## A3.1 AUTHORISATION.

| NAME | TITLE |
|---|---|
| David V Ford | Co-Director / Professor SAIL Programme |
| | |

## A3.2 DISTRIBUTION.

| NAME | TITLE |
|---|---|
| David V Ford | Co-Director / Professor SAIL Programme |
| All SAIL Staff | (Available via Confluence and OwnCloud). |

## A3.3 DOCUMENT HISTORY.

| VERSION | DATE | AUTHOR | DESCRIPTION | APPROVED BY |
|---|---|---|---|---|
| 1.1 | 18/12/2019 | Dan Thayer | First Version | Sharon Heys |
| 1.2 | 08/12/2020 | Rob Garlick | Annual review, updated as follows:<br>i). Correction of the policy reference number to SAIL-POL-024. The number was incorrectly assigned at the time of the original policy launch. | Sharon Heys |
| 1.3 | 16/03/2022 | Rob Garlick | Annual review, updated as follows:<br>i). Updated policy owner to **Dan Thayer.**<br>ii). Updated policy classification to **OPEN.**<br>iii). Addition of Compliance subheading and update of compliance text (in line with other policies). | Sharon Heys<br>Dan Thayer |
| 1.4 | 25/10/2022 | Rob Garlick | Update following HD-32578:<br>- Clarification update regarding file format for graphical data – no embedded or linked data.<br>- Emphasised requirement to remove embedded files, macros, comments, etc. | Rob Garlick<br>Hannah Davies |
| 1.5 | 05/04/23 | Hawys Rees | Annual review, updated as follows:<br>- Updated title to Output Policy Review.<br>- Added detail to the Kaplan-Meier survival plots graph.<br>- Updated the file format section to include formulas. | Dan Thayer |
| 1.6 | 30/05/2024 | Rob Garlick | Annual review, no updates to raise. | Dan Thayer |
| 1.7 | 06/06/2025 | Ashley Akbari | Annual review with updates to include Tier review model, change of minimum count number, expected conduct and clarifying certain points. | SAIL Operations group |