SWANSEA UNIVERSITY MEDICAL SCHOOL

OUTPUT REVIEW POLICY

Version	1.2 (RELEASE)
Issued By	SHARON HEYS
Classification	INTERNAL
Review Date	08/12/2020
Status	APPROVED
Reference	SAIL-POL-24
Note	UNCONTROLLED WHEN PRINTED

CONTENTS.

PURPOSE.

This document will detail the SAIL Output Review Policy which was defined and deployed as part of the Health Informatics Group & SAIL Senior Management approved and authorised ISMS project. This policy should be used in association with the SAIL Output Review Checklist and the SAIL procedure for Output Review Training.

SCOPE.

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

1. - OVERVIEW.

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

Our goal in implementing this process is to maximize 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).
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 2016 (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 information it holds is adequately de identified during its lifecycle. Data that is adequately anonymised 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 ensuring that anonymisation is maintained and upheld and to minimise any risks to the individual and to 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 judgement in order to be satisfied that outputs are safe.

This policy applies to standard SAIL projects where anonymization 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). Rules of thumb described in this handbook will be 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 datasets, date and geographic coverages within datasets, specified restricted data items with datasets, 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 Gateway, with only results requested out.
The expected operating model of SAIL is that all data preparation and analysis will be conducted within the SAIL gateway secure environment, 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, if there is a need for review by individuals who do not have access to SAIL, as long as they meet the requirements for safe outputs. However, researchers are advised to conduct review of intermediate results within the SAIL gateway whenever possible, as it minimizes risk, is more convenient for researchers, and reduces reviewer workload.

Individual-level data is not allowed out.
No individual-level person-based results are allowed out for anonymised research studies. This is true even if the results can be shown to be non-disclosive.

Cells shall not contain a value less than 5.
In accordance with the UK Information Commissioners 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 dataset. Normally, tabular and other results showing counts of less than 5 individuals or events related to individuals will not be allowed in accordance with Information Commissioner's Office (ICO) guidance. Researchers may choose to mask these values (replacing the exact value with "<5", for example) or deal with the situation in other ways, such as further aggregating categories.

Cases where a count of less than 5 may be derived are also disallowed. For example, if a table reports only one value less than five, 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 <5 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 <5 cases is just as useful as reporting that there were 3 cases (for example). Thus, in the spirit of minimizing outputs, as well as limiting risk and reviewer workload, exceptions are limited to those cases where there is a compelling reason for release.

Reporting of errors or missing data, in cases where it does not plausibly relate to a recognizable characteristic of a person, is not subject to the <5 rule.

Zero is disallowed where there is potential for disclosure.
Reporting of a count of zero is disallowed where, in combination with other information being reported, at has potential to disclose something meaningful about an individual or small group of individuals. In most situations this is not an issue.

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 number being reported. For example, if a bar chart reports a value <5, and the scale of the chart is low enough that it is possible to determine the exact number by measuring it, 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 have individual points for outliers, are typically disallowed, because they are showing individual-level data.

Kaplan-Meier survival plots often feature drops in the line that relate to small numbers of individuals, or a single individual. However, they also feature individuals being lost to follow-up over time, such that the

makeup of the sample at any given time is typically not known. Therefore, these types of plot often represent no disclosure risk, even if one might surmise that a drop in the plot relates to a small number of individuals. Disclosure potential of these plots will be considered by reviewers on a case-by-case basis.

Nothing that could be used for performance tracking of individual organisations may be released.
The data sharing agreements with data providers to SAIL disallow use of SAIL for 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 area are only allowed when explicitly mentioned in a project's IGRP application which has received approval.

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 a dataset that they own. 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 datasets, including one with a special requirement from the data provider, this requirement will apply to the whole result.

SAIL will maintain a list of datasets with special review requirements for reviewers to use and make this known to SAIL project applicants and users so they are also aware of these requirements before application and use on 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, track changes; etc.

Reviewers may request that outputs be resubmitted in another file format if hidden information (or the possibility of hidden 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 out where necessary or beneficial to publish or share outside of the secure environment. 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 responsibility of the researcher 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 researcher also has the responsibility to demonstrate that the outputs are safe. This means providing enough documentation for reviewers to understand the outputs, as well as answering questions or providing additional information where needed to demonstrate the safety of the results. Repeated 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 right 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 is a breach of the SAIL Data Access Agreement and will lead to researchers being disciplined.

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 make it clear what datasets were used within the outputs. Each row and column in a table, as well as graphs, should be clearly labelled. The reviewers should have enough information to understand the methods, such as statistical tests used, as well as any information relevant to potential disclosure (whether the results are from a sample or whole population, whether categories are exclusive and a cell value could therefore be derived from a total, etc).

All requests should include confirmation that the results are in line with the <5 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.

Minimizing Requests

Researchers should seek to request out only the outputs needed, as few times as possible. Sometimes corrections or updates to analysis are inevitable. However, repeatedly requesting the same or similar 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.

4. STANDARDS FOR REVIEWERS

Expectations

It is the responsibility of the reviewer to only release results that they understand and have confidence that they are not disclosive.

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 should be given for rejection. 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 only released if both reviewers 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 only be released 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 asks for it to be reconsidered, an additional senior reviewer will consider the request. As 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

Normally, an initial review of all outputs will be conducted within two working days. The result of this review will be a decision, or a request for clarification if additional information is needed.

Exceptionally large requests, as well as requests that contain complex or novel methods that require additional time to understand, may not be feasible within the two working day window. In these cases, a reviewer will inform the researcher that additional time is needed within two working days. If the researcher expects that a request may be challenging, giving advanced notice to the reviewers may be useful.

Any person, subject to this policy, who fails to comply with the provisions as set out above or any amendment thereto, shall be subjected to appropriate disciplinary or legal action in accordance with the Swansea University, College of Medicine & Health Informatics Group Disciplinary Code and Procedures.

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

A. DOCUMENT MANAGEMENT.

A.1 AUTHORISATION.

NAME	TITLE
David V Ford	Co-Director / Professor SAIL Programme
Sharon Heys	Head of Legislation and Due Diligence

A.2 DISTRIBUTION.

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

A.3 REFERENCES.

DOCUMENT
Griffiths, E., Greci, C., Kotrotsios, Y., Parker, S., Scott, J., Welpton, R., … Woods, C. (2019). *Handbook on Statistical Disclosure Control for Outputs*. https://doi.org/10.6084/M9.FIGSHARE.9958520.V1
Ritchie, F., & Elliot, M. (2015). Principles- versus rules-based output statistical disclosure control in remote access environments. *Working Papers*. Retrieved from https://ideas.repec.org/p/uw

A.4 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: i). Correction of policy reference number to SAIL-POL-024. Number was incorrectly assigned at original policy launch	Sharon Heys