Medicines & Healthcare products
Regulatory Agency

CPRD Safe Outputs Guidance

Published 15 August 2024

Contents

Introduction 3
CPRD Approach to Releasing Outputs 4
Researcher Responsibilities 5
Automated Checking of Research Outputs (ACRO) 6
Self-Declaration 7
Safe Statistics 10
Unsafe Statistics 12
Exceptions 14
Other Methodologies 14
Appeals 14
Statistical Disclosure Control Options 15
Recommended SDC Resources 16

Introduction

CPRD collects pseudonymised patient data from a network of GP practices across the UK. Primary care data are linked to a range of other health related data to provide a longitudinal, representative UK population dataset. The CPRD Safe Trusted Research Environment (TRE) gives approved researchers with approved projects access to anonymised CPRD data within a highly secure computing environment. In line with the Five Safes framework, CPRD has developed an "Airlock" which uses both human and automated checking to ensure any outputs leaving CPRD Safe are "Safe Outputs". Checking of research outputs is a key part of the controls CPRD uses to ensure safe legal use of anonymised data for research. Output checking has two main goals:

- To minimise the risk of potentially disclosive results – meaning results with the potential to reveal information about an identifiable individual or GP practice.

- To ensure that the data is used only in line with CPRD policies, agreements with data providers, and the Research Data Governance (RDG) approval that is in place for use of the data.

The objective in implementing output checking is to maximise the utility of the data for research, while safeguarding CPRDs ongoing operations. The risks output checking seeks to manage are:

- Breaching individuals' privacy – revealing information about a known individual (spontaneous recognition) or revealing information from which an individual may become identifiable via deductive disclosure.

- Legal violations – CPRD's legal basis for operation is that CPRD's safeguards and processes ensure the anonymity of the data made available for approved observational research. If CPRD are found to release potentially disclosive results (even if no individual is identified), CPRD may be in breach of the law and/or regulatory or ethical approvals.

- Reputational damage – CPRD relies on the trust and cooperation of data providers (including patients and the public) and other stakeholders to operate. If CPRD processes are perceived to be unsafe, then CPRD risk not receiving data and being unable to operate.

This document provides guidance to researchers on preparing outputs prior to submitting them for review and approval for export through the CPRD Safe Airlock.

CPRD Approach to Releasing Outputs

CPRD operates a default-open risk assessment approach, and releases all outputs unless there are demonstrably meaningful risks in doing so. This is consistent with the aim of output checking (to maximise the utility of the data subject to addressing disclosure risks) and reflects that CPRD research outputs are very low risk by their nature.

CPRD operates a principles-based output checking service. Rather than having hard rules, CPRD uses a set of principles, and relies on trained and experienced output checkers applying these and their judgement to be satisfied that the outputs are safe. Exceptions can be permitted subject to a researcher demonstrating that their output is non-disclosive, the detail in the output is important, and their request for an exception is a rare occurrence.

Outputs are processed promptly, with the timeline directly related to the clarity of the output request and the use of Automated Checking of Research Outputs tools (see below), where appropriate. Well justified and clearly explained outputs, following CPRD guidance, are more straightforward to assess. More complex or less well justified output requests will require additional time for thorough review.

Researcher Responsibilities

It is the responsibility of the researcher to produce Safe Outputs. CPRD have created guidance and training, including this document, for researchers on what exports are possible, and how to appropriately summarise and minimise data.

Data must be processed within CPRD Safe, with only results submitted for output checking. Outputs must be submitted for output as a single file (.zip files are permitted to enable meta-data and multiple files to be submitted as a single request). The expected operating model of CPRD is that all data preparation and analysis will be conducted within CPRD Safe, with only results requested and (where meeting CPRD criteria) approved for output.

Outputs are not restricted to final results. Intermediate results are allowable, if there is a need for review by individuals who do not have access to CPRD Safe, and they meet the requirements for Safe Outputs. However, review of intermediate results should be conducted within CPRD Safe whenever possible, as it minimises risk, is more convenient for researchers, and reduces output checker workload.

Please note that screen sharing when using CPRD Safe is only permitted under very limited circumstances. Screen sharing within a study team (e.g. among collaborators on an RDG approved protocol) is permitted only when individuals logging into the workspace would be impractical. Screen sharing within an organisation (but outside of the study team) is only permissible (i) on a clear need-to-know basis, and (ii) when it would be impractical to onboard the individuals to the workspace via CPRD processes. Video calls including screen-sharing must only be conducted using enterprise editions of video calling software managed by the researcher's organisation, and recording must never take place. Similarly, screen shots or screen grabs of CPRD Safe are never permitted.

Public screen-sharing should never take place. All possible precautions should be taken to ensure researchers physical space remains secure, and access to CPRD data remains confidential.

Any deliberate attempt to circumvent the CPRD Safe Airlock mechanism or the output checking system is a breach of CPRD contractual controls and will lead to CPRD intervention. Use of the output checking system will be audited in accordance with standard CPRD procedures. The consequences of inaccurate or inappropriate use may include loss of data access (including for ongoing or future research projects), increased scrutiny and longer timelines for output checking, a reduction in other services included under the licence agreement, and/or legal action.

If researchers are in any doubt as to their responsibilities, they should contact enquiries@cprd.com.

Automated Checking of Research Outputs (ACRO)

CPRD strongly recommends the use of the ACRO tool when conducting analysis within CPRD Safe. ACRO tools can be applied to certain analyses conducted in Python, R, and STATA, but must be applied when the analyses are undertaken – not as an add on afterwards.

ACRO is an open-source tool for automating the statistical disclosure control (SDC) of research outputs. It assists both researchers and output checkers by distinguishing between research outputs that are safe to release, outputs that require further analysis, and outputs that cannot be released because of substantial disclosure risk.

It does this by providing a lightweight "skin" that sits over well-known analysis tools, in a variety of languages researchers might use. This adds functionality to:

- identify potentially disclosive outputs against a range of commonly used disclosure tests,
- suppress outputs where required,
- report reasons for suppression,
- produce simple summary documents that CPRD output checkers can use to streamline their work.

The use of ACRO will reduce the time required for output checking. Non-disclosive outputs created using ACRO, where applicable, will be approved more quickly than those where ACRO has not been used.

More information on ACRO can be found on the ACRO GitHub site. Training videos can be found on the ACRO YouTube site.

Self-Declaration

CPRD encourages researchers to develop their outputs to a fixed set of rules to allow for efficient output checking. Only files that the researcher has already checked and believes to be safe and in line with CPRD policies and RDG approval should be submitted for output checking. The rules form the first set of questions on the Output Declaration Form (ODF) within CPRD Safe. If the following rules are not met, the request will automatically be rejected.

Requested outputs are consistent with the RDG approved protocol associated with this workspace.

Where outputs are requested under an RDG-approved project, data used to produce the results must be limited to what is approved for the project (including specified datasets, dates, and geographic coverage within datasets, specified restricted data items within datasets, etc.). The analysis must align with those described in the application, and any other specific restrictions to the output described in the RDG protocol must be adhered to.

No event or patient level data are included in the requested outputs.

Individual-level data will not be approved for release from CPRD Safe. No individual-level person-based results will be approved for release for anonymised research studies. This is true even if the results can be shown to be non-disclosive.

All requested outputs are sufficiently clear and comprehensible to permit output checking without the need for dataset- or project-specific knowledge.

The researcher has the responsibility to provide enough documentation for output checkers to understand the outputs, as well as answering questions or providing additional information where needed to demonstrate the safety of the results. Submission of outputs without adequate documentation will lead to immediate rejection and return to the researcher.

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 output checkers.

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 output checkers should have enough information to understand the methods, such as statistical tests used, as well as any information relevant to potential disclosure.

All requested outputs are static.

Graphs should be output as fixed images (e.g. .png), rather than with the data used to generate them sitting behind the graph.

All requested outputs use permitted file types.

All outputs must use one of the following file types. When multiple files are being submitted in one output request, they must be zipped (.zip).

- Generic: .bmp, .csv, .gif, .gz, .jpeg, .jpg, .log, .notebook, .pdf, .png, .sql, .svg, .TeX, .tsv, .txt, .zip
- CPRD Code Browser: .pmb, .ppb
- Excel: .xls, .xlsm, .xlsx
- JavaScript: .json
- Jupyter: .ipynb
- Markdown: .md, .rmd
- Python: .py, .py3, .rst
- R: .R, .Rda, .Rdata, .Rds, .tar, .tar.gz
- STATA: .ado, .do, .dta, .gph, .mata, .pkg, .smcl, .sthlp, .toc

No hidden information has been included (e.g., embedded files, comments, track changes).

Hidden results or other hidden information (even if not results) is not allowed (e.g., hidden columns/rows/sheets in an Excel spreadsheet, links to external files in Excel formulae, linked data used to generate pivot tables, embedded Excel files within an Office document, comments, track changes).

Safe Statistics

Researchers will be asked via the ODF to self-declare whether their outputs include “safe” or “unsafe” statistics. A safe statistic has no meaningful disclosure risk in most circumstances. Safe statistics, and their criteria, include:

- Statistical hypothesis tests (e.g., t-test, chi-square, R-square, standard errors)
    - Run on a minimum of five patients
- Coefficients of association (e.g., estimated coefficients, models, AN(C)OVA, correlation tables, density plots, kernal density plots)
    - Residual degrees of freedom (number of observations less number of variables) exceeds five
    - The model is not saturated (i.e., not all variables are categorical and fully interacted)
    - Outputs do not include a regression with a single binary explanatory variable
- Shape (e.g., standard deviation, skewness, kurtosis)
    - Any standard deviations are greater than zero
    - All statistics of shape were calculated for a minimum of five patients or GP practices
- Mode
    - mode is not the only value (i.e., different observations have different values)
- Non-linear concentration ratios (e.g., Herfindahl-Hirchsmann index, diversity index)
    - N>2
    - H<0.81
- Gini coefficients or Lorenz curves
    - N>2
    - The coefficient is less than 100%

Exceptions may be possible for modes and non-linear concentration ratios that do not meet these criteria (see below). If any other criteria are not met, the output request will be automatically rejected.

A response to submissions that only include safe statistics should be provided within three working days of submission. It is recommended to separate safe statistics from “unsafe” statistics to enable the most efficient output checking.

Unsafe Statistics

An “unsafe” statistic has meaningful disclosure risks, and so those risks need to be assessed on a case-by-case basis before the output can be approved for release. Unsafe statistics, and their criteria, include:

- Frequencies (e.g., frequency tables, histograms, shares, alluvial flow graphs, heat maps, line graphs, pie charts, scatter graphs, scatter plots, smoothed histograms, waterfall charts)
    - All counts <5 and frequencies derived from groups containing <5 patients or GP practices have been suppressed
    - All zeroes and full cells (100%) are evidential or structural (i.e., something you would expect)
    - Underlying values are genuinely independent (i.e., they do not come from the same patient, the patients do not all have the same family number and do not all come from the same GP practice)
    - The categories are comprehensive and apply to all data (i.e, all categories of each categorical variable are presented).
- Position (e.g., median, percentiles, box plots)
    - The numbers for each group (and complementary groups) are ≥5
- Extreme values (e.g., maxima, minima)
    - The maximum or minimum presented are non-informative and structural
- Linear aggregates (e.g. means, totals, simple indexes, linear correlation ratios, bar graphs, mean plots)
    - The linear aggregates have been derived from groups containing ≥5 patients or GP practices
    - The P-ratio dominance rule has been calculated and is greater than 10% (ACRO will check this automatically)
    - The N-K dominance rule has been calculated for the 2 largest values and is less than 90% (ACRO will check this automatically)

- Odds ratios, risk ratios, or other proportionate risks
    - The underlying contingency table has been produced and is included in the requested outputs
- Hazard and survival tables (e.g., tables of survival/death rates, Kaplan-Meier graphs)
    - The number of patients who survived is ≥5
    - Exit dates are relative, not absolute
    - There are no dates with a single exit

Exceptions may be possible for unsafe statistics that do not meet these criteria (see below).

A response to submissions that include unsafe statistics should be provided within five working days of submission, dependent on ACRO tools being applied where applicable. Where ACRO has not been used, a response should be provided within fifteen working days. It is recommended to separate statistics which have been produced using ACRO from those which have not to enable the most efficient output checking.

Exceptions

If a researcher is in any doubt that their requested outputs fully meet the rules and principles detailed in the ODF, they will be directed to complete a self-declaration of exceptions in CPRD Safe. This includes questions to prompt development of a risk mitigation strategy. This provides a route for exceptions where principles-based disclosure control will be applied, which will always include review by the CPRD Information Governance team on a case-by-case basis. Requests for exceptions should be rare occurrences.

All requests must include justification that shows the output is safe and important to release if the researcher is requesting an exception. The risk mitigation strategy will be reviewed to assess the reason the exception is required (including public health benefit), to ensure the outputs specifically relate to an RDG approved study, and the plans for onwards use.

A response to submissions that include exceptions should be provided within fifteen working days of submission. This does not include the time taken by the researcher to respond to any requests from the output checkers.

Other Methodologies

A wide range of methods are used in approved research using CPRD data, not all of which are covered by this guidance. Rare and novel methodologies will be handled as exceptions. Researchers are encouraged to discuss such methods with CPRD output checkers in advance of submitting an output request.

Appeals

Outputs will be automatically rejected if they do not meet CPRD disclosure control rules, and by output checkers if they are assessed to be potentially disclosive. However, if a researcher believes their requested outputs are important to be released and there has been a failure of process, they should contact enquries@cprd.com within 10 working days of rejection with their justification. Researchers cannot appeal on the basis they simply disagree with the decision. CPRD may be able to release the requested outputs under the principles-based approach to output checking with adequate justification.

If a researcher appeals, the request will be escalated to the next level of output checking seniority for consideration. The original and senior output checkers will discuss the issue, and the decision changed only if all are satisfied that the outputs can be released.

Statistical Disclosure Control Options

To remove or reduce the risks associated with research outputs, a range of statistical disclosure control (SDC) options are available. CPRD do not require or recommend that a specific SDC option is used – not all approaches are appropriate for all outputs, and SDC options offer flexibility depending on the message the results are aiming to convey. However, some common options are presented below.

Limiting source data combinations

Only permitting specific tabulations of variables which allow each cell to have sufficient observations to prevent disclosure risk. This often involves collapsing or merging categories to increase observation counts. This method increases the likelihood of the same changes being applied across multiple tables, reducing differencing risk.

Cell suppression, with adjustment of totals

Removing cells with fall below the threshold, usually replacing with blanks or some other non-informative filler. It requires additional suppression or recalculation of totals to prevent disclosure by subtraction.

Cell suppression, with secondary suppression

Removing cells which fall below the threshold, along with other cells above the threshold to protect the hidden cells. This maintains marginal totals, although at the cost of missing data in specific cells. This may be appropriate where consistency in totals across tables is valued.

Noise addition – simple

Tables are adjusted by adding a small amount of random noise so that the true value in the cell is uncertain. Noise can be additive (x becomes x+y, where y is a random value -n to +n), scaled (x becomes x+my, where m grows as x increases) or multiplicative (x becomes xy). The values associated with the cells may need to be adjusted as well. Because noise addition is not obvious by looking at the table, it must be clearly highlighted in the methodological notes to the table. Some values will be unaffected by the noise, as occasionally adding zero noise should be one of the permissible outcomes for a random noise allocation process.

Noise addition – differential privacy

Considers what values could have been in the dataset, not what were, and then adding noise. The noise – the random value that is added – ensures that no individual's inclusion or exclusion from the dataset can significantly affect the results of analyses.

Rounding – conventional

Adjusting the values in all cells in a table to a specified base to create uncertainty about the real value for any cell. The larger the base rounding value, the more protection is provided, although more accuracy is lost. For additional protection, rounding can be carried out to multiples of the rounding value (e.g. round to nearest 5 in 80% of cases, round to nearest 10 in 15% of cases, round to nearest 15 in 5% of cases).

Rounding – controlled

Rounding using linear programming techniques to round cell values up or down by small amounts.

Recommended SDC Resources

- The SACRO Guide to Statistical Output Checking, 2023, was developed as an output of the project SACRO (Semi-Automatic Checking of Research Outputs) funded by UK Research and Innovation. It discusses the operational and statistical theory underlying output SDC, lists the rules-of-thumb to be followed for outputs (organised by class), and provides frequently asked questions on both statistics and outputs.

- The SACRO Report, 2023, outlines the project to automate checking of most common statistics, support researchers using the major analytical languages, and support secure environments through a process of co-deign.

- The Handbook on Statistical Disclosure Control for Outputs, 2019, was produced by the Safe Data Access Professionals Working Group and explains methods to apply SDC for a wide range of outputs.

- The Five Safes Framework, 2017, explains the Five Safes through several practical examples and case studies.