Semi-Automated Checking of Research Outputs (SACRO)

DARE Driver Project Final Report

Principal Investigator J.E. Smith (University of the West of England).

Contributors to this report: M. Albashir, S. Bacon, B. Butler-Cole, J. Caldwell, C. Cole, A. Crespi Boixander, E. Green, E. Jefferson, Y. Jones, S. Krueger, J. Liley, A. McNeill, K. O'Sullivan, K. Oldfield, R. Preen, F. Ritchie, L. Robinson, S. Rogers, P. Stokes, A. Tilbrook, P. White.

UK Research and Innovation

HDRUK Health Data Research UK

ADRUK Data-driven change

1. Executive Lay Summary

This project aimed to address a major bottleneck in conducting research on confidential data - the final stage of "Output Statistical Disclosure Control" (OSDC). This is where staff in a Trusted Research Environment (TRE) conduct manual checks to ensure that things a researcher wishes to take out - such as tables, plots, statistical and/or AI models- do not cause risk to any individual's privacy. To tackle this bottleneck, we proposed to:

- Produce a consolidated framework with a rigorous statistical basis that provides guidance for TREs to agree consistent, standard processes to assist in Quality Assurance.
- Design and implement a semi-automated system for checks on common research outputs, with increasing levels of support for other types such as AI.
- Work with a range of different types of TRE in different sectors and organisations to ensure wide applicability.
- Work with public and patients to explore what is needed for public trust, e.g., that any automation is acting as "an extra pair of eyes": supporting not supplanting TRE staff.

Supported by funding from DARE UK (Data and Analytics Research Environments UK), we met these aims through production of documentation, open-source code repositories, and a 'Consensus' statement embodying principles organisations should uphold when deploying any sort of automated disclosure control.

Looking forward, we are now ready for extensive user testing and refinement of the resources produced. Following a series of presentations to national and international audiences, a range of different organisations are in the process of trialling the SACRO toolkits. We are delighted that DARE UK has awarded funding to support a Community of Interest group (CoI). This will address ongoing support and the user-led creation of 'soft' resources (such as user guides, 'help desks', and mentoring schemes) to remove blocks to adoption: both for TREs, and crucially for researchers.

There are two other areas where we are now ready to make significant advances: applying SACRO to allow principles-based OSDC for 'conceptual data spaces (e.g. via data pooling or federated analytics) and expanding the scope of risk assessment of AI/Machine Learning models to more complex models and types of data.

2. Description of Project Outputs

A high-level description of the project's intended outputs, and how we organised to create them, is described in the project proposal. For ease, this is available online[1], alongside various documents comprising or summarising the outputs from different work packages. In the following subsections we describe the specific outputs produced by each work package, beginning with the conceptual framework, followed by the technical work packages, then those involving external stakeholders. Full details of feedback from all the TREs are provided on the sharepoint site, and for illustration, we provide a case-study from the Grampian Safe Haven.

2.1. Conceptual Framework

The main aim here was to deliver a comprehensive new guide to output statistical disclosure control (OSDC), including the latest developments in both theory and practice. These developments had mostly been disseminated through piecewise and informal communication, creating a need for a clear authoritative foundation. The specific objective was to produce the formal document, developed in consultation with the SACRO network and others, as well as additional web tools. This goal has been achieved. Outputs include:

[1] Via UWE sharepoint. here

- The main document covering theory and practice[2], using a new taxonomic model which underpins SACRO
- A structured document allowing all catalogued statistical outputs to be grouped and searched, to aid output checkers in making use of the new guidelines
- Coding for HTML 'popups', designed to be incorporated into SACRO and TRE support environments

The guide formalises a radical new approach to output checking which views output risks as being associated not with a particular statistic, but with a class of statistics. For example, frequency tables and pie charts have the same disclosure characteristics. Output checking then moves from being "what are the rules associated with this output?" to "What type of output is this?" This taxonomic approach (the 'stat barn') is directly implemented in SACRO – it is how it manages an almost infinite range of statistical types with a finite set of operating rules.
The guide and lookup table have been circulated amongst the SACRO network; the draft final versions are being circulated to wider, international, networks (including statistical agencies, banks, and data archives). A paper describing the approach was presented at the biannual UNECE expert meeting on statistical data confidentiality[3], and has been circulated to our network (e.g., UNECE participants, SDAP) for further comments, as well as to wider SDC groups. The coding of the text describing risks and mitigation strategies for specific outputs is complete, but it was decided design issues around implementation should be led by the CoI group.
Lessons learned: developing a comprehensive, evidenced, structured, authoritative guide which also met the needs of more casual users proved too much of a task. We concentrated on the former and are collaborating with the SDAP group to update their popular, more informal, examples-based manual.
Future plans: We expect to publish one theory and one applied journal article in 2024. The Guide will also be evaluated by the CoI group to ensure (a) buy-in to the theoretical developments, and (b) incorporation of 'best practice operational elements from the CoI members. It will form the centrepiece of an international workshop to be organised by the Bundesbank in 2024, and of an ESRC experimental 'data academy'.

2.2. The ACRO engine

This focussed on developing the key under-pinning "ACRO engine", which conceptually is composed of two parts. The first is a Python package "acro" that leverages industry-standard tools[4] to perform analytic queries, and simultaneously conducts and reports on OSDC checks appropriate to the analysis type, informed by a simple human and machine-readable file specifying the TRE's risk appetite for a given dataset. This package is available from PyPI via `pip install acro`. The second part is a set of interfaces in Python, R and Stata that researchers can use as drop-in replacements for standard analytic commands. For example, to create tables, in Python the standard Pandas command `pd.crosstab()` becomes `acro.crosstab()`; in R, `table()` becomes `acro_table()`; and in Stata `table` becomes `acro table`.

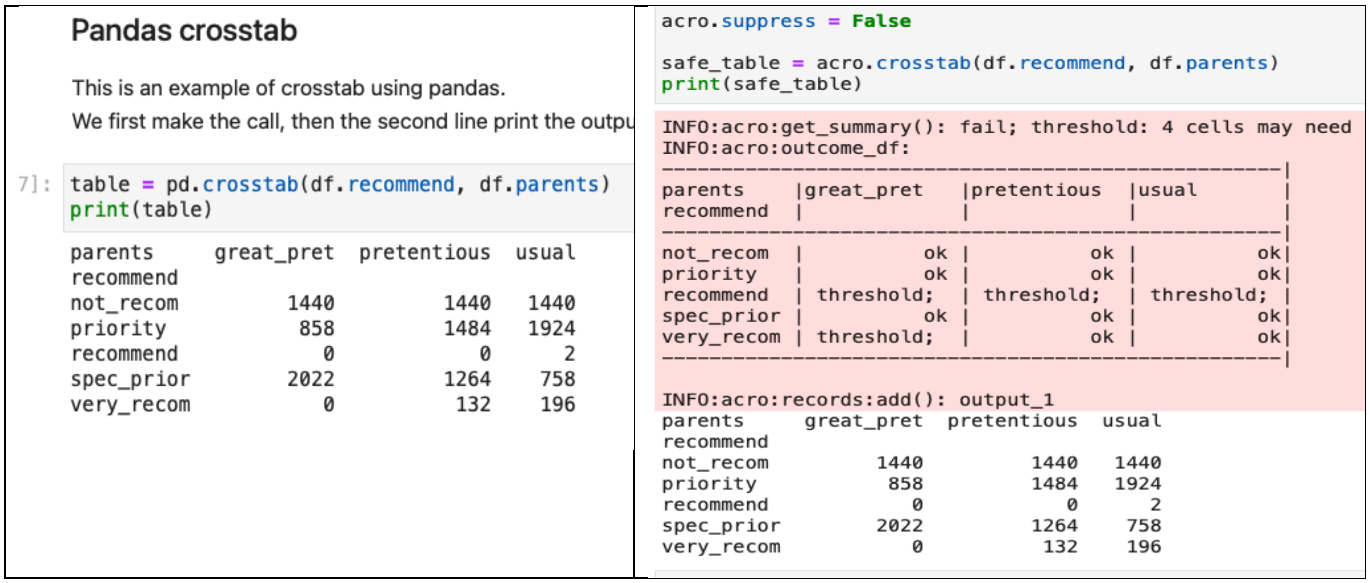
As Internet access is restricted from within TRE (virtual) environments, the acro package has extensive docstrings to support Python's inbuilt help mechanisms, e.g., `help(acro.ACRO.command)`. A separate file (currently acro.R – but see below) provides interfaces to R for all analytic functions, and `acro_help(command)` provides pass-through access to Python help commands. The Stata interface currently supports versions <16. Transfer of data and queries from Stata/R to the Python acro package, and analytic/disclosure results in the other direction, is all handled in memory via the packages SFIToolkit (Stata) and Reticulate (R), enabling a single source of truth for OSDC that is covered by a comprehensive set of unit tests. In response to feedback from TRE's, at the time of

[2] Ritchie, F., Green, E., Smith, J., Tilbrook, A., & White, P. (2023). The SACRO guide to statistical output checking (Version 1). Zenodo. https://doi.org/10.5281/zenodo.10054629
[3] Derrick, B., Green, E., Richie, F., and White, P.: Towards a comprehensive theory and practice of output SDC. United Nations Economic Commission For Europe Conference Of European Statisticians, Expert Meeting on Statistical Data Confidentiality, September 2023, Wiesbaden. Online at https://unece.org/sites/default/files/2023-08/SDC2023_S5_2_UWE_Ritchie_D.pdf.
[4] Pandas (https://pandas.pydata.org) and statsmodels (https://www.statsmodels.org)

writing we are finalising submission of an R package, which will permit researchers to install the R interface, and underpinning acro suite in any TRE that supports access to CRAN[5] (or a copy thereof).

TREs were involved in co-design of these tools from the outset. A series of "reverse science cafés" was used to understand operational issues that would affect the design and testing process – for example, not all TREs support the use of Python in their 'main' environment. Thereafter, rounds of iterative development and testing were used, with a combination of regular 1:1 meetings between individual TREs and developers, and two hybrid meetings (June and October) with sessions dedicated to discussion, feedback, and prioritising issues. Throughout this the technical teams developing the acro library were in constant communication with the developers of the viewer tool that ingests its outputs to design the structure of the JSON file which acts as the interface between a researcher's 'acro session' and the TRE viewer, or can be transmitted in federated analytic settings.

Table 1: below shows a side-by-side comparison of the researcher view with and without using acro.



Pandas crosstab

This is an example of crosstab using pandas.

We first make the call, then the second line print the outpu

```
7]: table = pd.crosstab(df.recommend, df.parents)
    print(table)
```

parents	great_pret	pretentious	usual
recommend			
not_recom	1440	1440	1440
priority	858	1484	1924
recommend	0	0	2
spec_prior	2022	1264	758
very_recom	0	132	196

```
acro.suppress = False

safe_table = acro.crosstab(df.recommend, df.parents)
print(safe_table)
```

```
INFO:acro:get_summary(): fail; threshold: 4 cells may need
INFO:acro:outcome_df:
-----------------------------------------------------|
parents     |great_pret    |pretentious   |usual       |
recommend   |              |              |            |
-----------------------------------------------------|
not_recom   |           ok |           ok |          ok|
priority    |           ok |           ok |          ok|
recommend   | threshold;   | threshold;   | threshold; |
spec_prior  |           ok |           ok |          ok|
very_recom  | threshold;   |           ok |          ok|
-----------------------------------------------------|

INFO:acro:records:add(): output_1
```

parents	great_pret	pretentious	usual
recommend			
not_recom	1440	1440	1440
priority	858	1484	1924
recommend	0	0	2
spec_prior	2022	1264	758
very_recom	0	132	196

Table 1: Illustration of researcher views for equivalent commands in pandas (left) and acro (right). Output in pink also forms part of a report created for TRE staff. This table fails the check that outputs should not report on groups containing fewer than a set level of respondents.

The main ACRO package now provides support for researchers to conduct analysis using a range of regression statistics, survival analysis (a particular request of TREs), and hierarchical tables with various aggregation functions (e.g., counts, sums, statistical descriptors) as defined and 'approved' by the new guide. Plots such as histograms are also supported, since the taxonomy allows these to be checked using existing checks. Fuller details may be found via the project documentation on GitHub, or in the interim paper[6]. To assist the 'principles-based approach' to OSDC, outputs can be produced with or without automated suppression – for example, blanking disclosive cells and recalculating marginal totals in a table.

A set of 'session management commands' enhance the researcher experience. For example, researchers can list outputs, remove those they have been warned are disclosive (and have hopefully reworked), rename outputs, and add 'unsupported outputs'. The latter could be outputs from unusual forms of analyses, text files, or their code. To aid researcher/output checker communication and audit, researchers can add comments, and exception requests to any output in a 'session', and the latter are requested again when the researcher 'finalises' an acro session containing outputs flagged as disclosive.

[5] https://cran.r-project.org

[6] Smith, J., Preen, R., Albashir, M., Ritchie, F., Green, E., Davy, S. & Bacon, S. "SACRO: Semi-Automated Checking Of Research Outputs". Presented at UNECE Expert meeting on Statistical Data Confidentiality, Wiesbaden, Sept 2023. Available here

The code repository, documentation, coding standards/reports are available on GitHub[7]. The R and Python code has been tested with open source datasets by the developers, and in different TRE environments on open source and 'confidential' datasets. The Stata code has been tested on some open source datasets, but less extensively beyond that due to (i) a smaller user base within the project, and (ii) a major change in the syntax for v16.

As part of our outreach, we are now in the process of helping TREs beyond the project consortium to install and test ACRO. We are also working closely with Public Health Scotland's team who are specifying the 'National SafeHaven 2.0' (to be delivered by EPCC early 2024), providing documentation to support IT Governance /software risk analysis of the SACRO software. We will make this documentation available to assist other TREs.

Future plans: Having successfully delivered the initial toolkits and libraries, work is now needed on a number of fronts to achieve the desired impact in terms of freeing up resources for TREs. We are delighted that DARE UK is supporting a CoI group, under whose auspices much of this work will progress, which includes:

- Creating community-designed resources for TREs and (especially) researchers to promote adoption.
- Ongoing development of both the ACRO library, and the 'translation functions' to enhance functionality, support more analyses and respond to inevitable issues as they arise from an increased volume of testing.
- Removing blocks to researcher adoption as they are identified. For example, extension to cover more of the R 'tidyverse' family (dplyer, etc.), rewriting parts of the translation to support different Stata versions.
- Integrating with other related code initiatives. For example, we have identified a range of modalities for integrating SACRO code within DataShield to either (i) extend their functionality to provide support for principles-based OSDC; or (ii) enforce suppression within SACRO and embed within DataShield to support Python-based analytics within their strict rules-based approach.

Beyond the CoI group, we are now in position to support a major piece of work to be done around integration with different models of data pooling and federated analysis as described in Section 4.

2.3. The SACRO viewer

The goals of this work package were to:

- Develop a desktop application to make output checking and release approval easier
- Incorporate the outputs of the ACRO library to further ease the output checking task
- Include TREs in the technology and user interface design choices to ensure utility of outputs.

We have met these goals. Feedback has been gathered from a mix of live demonstrations and 'hands-on' testing within TREs, although the former has dominated at the time of writing, due to the necessary complexity of getting software installed inside TREs. The developers collaborated closely with the ACRO library team, both in defining the functionality of the viewer and integrating the two systems. We've worked with SACRO's TRE partners to understand the constraints on TRE-installed software, which informed our technical design decisions.

The SACRO Outputs Viewer runs on Windows, Linux (Ubuntu/Debian), and macOS. It is packaged using standard tooling (MSI on Windows and .deb on Linux) and uses the Electron browser and is written in JavaScript and Python. The Windows/Linux installers[8] bundle all dependencies, including JavaScript and Python runtimes.

Figure 1 (below) illustrates the TRE output checker's view of the outputs created by the researcher in Table 1. As the testing has progressed, we are in the process of implementing requests from TREs – such as support for viewing researchers' code scripts with syntax highlighting, which had been viewed as a bottleneck.

[7] https://github.com/AI-SDC/ACRO, part of the AISDC organisation which also hosts the machine learning toolset

[8] Available at https://opensafely.org/sacro/latest-windows-build and https://opensafely.org/sacro/latest-linux-build

Future plans: Beyond on-going support and testing, the design of the viewer is considered to be largely complete, and feedback from TREs is that it is easy to use, suggesting it is not a priority for training resources. However, as the TRE infrastructure landscape evolves, and new lightweight dashboard frameworks such as Streamlit increase adoption, the question of choice of tools for implementing the viewer design inevitably remains open. We will also need to consider whether it is most appropriate to include outputs from checking Machine Learning models.

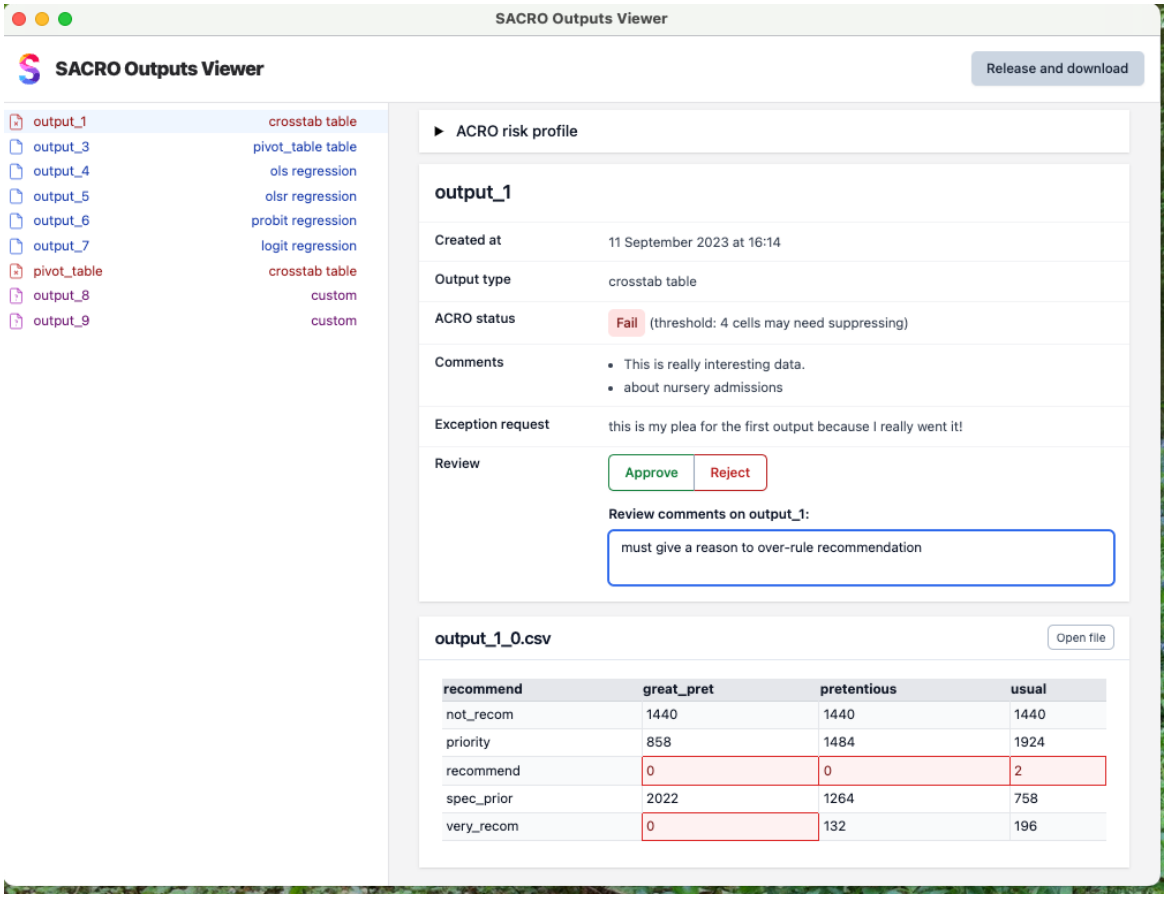


Figure 1: Illustration of output viewer for the analysis shown above

2.4. Automated Risk Assessment of Machine Learning Models

This work package built on the understanding and resources produced during the DARE UK GRAIMatter project, but also considered (i) rapid developments in the ML field during the project; and (ii) an increasing number and urgency of requests at TREs (inside and beyond SACRO) for risk analysis of trained ML models. In response to the latter, we adapted some of the project planning, identifying a number of 'user journeys' as per Figure 2 below, and then doing a major code refactor and creating new tools for risk assessment. These have been successfully trialled in TREs to assess 'live' cases of ML algorithms produced for egress. All tools are available on GitHub[9].

These tools are designed to be simple for TREs to run. They need only edit a simple text file to specify the 'user story' a researcher has followed, and the locations of their files, then call a script that automatically runs the most informative attacks available, gathers a range of metrics, and produces a risk report. Depending on the model and the 'user story' this risk report may provide sufficient 'reasonable grounds' for TRE staff to make a decision or provide evidence for more 'expert' advisors. Liaison with the team conducting PIE work helped inform the consensus statement around automated checking and the future use of AI.

[9] https://github.com/AI-SDC/AI-SDC

Case Study from a TRE perspective - Grampian Data Safe Haven:

Restrictions on Python in our 'main' safe haven meant that the Python/R install was on a segmented environment. Challenges to install were further compounded by versioning of Python, so we worked with the SACRO team to create a Docker that allowed us to run a higher version of Python to test ACRO. The team also modified the SACRO package to run on a lower version of Python, which allowed us to run the R version. Because of the environment limitations, we adopted two test methodologies – one, as output checkers running SACRO on projects, and two, testing as researchers. This allowed us to understand the tool from both sets of users. We then began testing on real project data (code and previously requested files for disclosure checks) of historic projects, first, as a means of testing on real data, and second, to establish benchmarks against human output checking. This allowed us to establish whether SACRO can be used as a first set of eyes in the disclosure-checking processes and whether it was as effective as humans at highlighting potential issues. In both cases, we found SACRO met the accuracy of a human output checker, and as we continue to test it against output requests, we will look to integrate SACRO fully in our output checking process. We have two suggestions for the SACRO team: (1) expand the tool to cover more R statistical analysis types (since the tool relies on Pandas and Statsmodels Python libraries, this proves limiting for more complex statistical analysis methods for researchers using R); (2) clearer documentation for minimum system requirements, package/library dependencies, installation process, and making SACRO available on R-Cran so that researchers are able to install it themselves directly since most TREs offer some mechanism for self-service (using a CRAN-mirror or PiPy-mirror, etc.).

We found generating output json files, interacting with the console to add comments and finalise outputs etc., to be intuitive and overall had no issues. The application worked well, installation was straightforward, and the GUI design was simple and professional. The GUI highlighted cells of concern and allows us to inspect these further to assess whether these were accepted values or ones that needed further explanation from researchers or were disclosive. The option to include output files not supported by the SACRO functions is a major plus: this still makes the SACRO module a useful vehicle for saving Researcher outputs with comments etc., ready for disclosure checking in an application where all of their outputs and comments can be viewed together.

A significant piece of work for this work package was to establish links between the way that ML researchers measure 'privacy leakage', typically via 'Membership Inference Attacks' (MIA), and disclosure control theory for 'traditional' analyses. Using a combination of think-pieces and workshops to establish a common language of discourse, we have implemented OSDC concepts such as threshold counts, class disclosure, and complementarity in the form of a range of 'structural' attacks on trained ML models to allow comparisons to be drawn across these different fields. We are currently analysing results from extensive experiments across a range of datasets and ML model types. We have shown that we can predict with high accuracy when a tree-based model is 'unnecessarily risky', i.e., when it is one of the most disclosive configurations for its type, and also that this metric is only weakly correlated with the model's accuracy on the task for which it was trained. A paper on this topic is under preparation for journal submission.

Future Plans: Within the CoI workgroup there are a number of activities identified by TREs around the provision of mechanisms for external support for TREs, and co-designing the 'risk reports' produced by the AI-SDC toolkits. Experience in supporting TREs suggests that it would be helpful to build work on 'user stories' into the project approval stage as well.

More widely, the ML field is in a state of flux when it comes to privacy risk assessment, with two 2023 papers demonstrating that the claims of MIA are frequently vastly overstated. We suggest this stems from (i) an overwhelming focus on deep neutral networks for image-based problems, and (ii) a lack of regard of the context –

for example, privacy risk in the 'average case' is of generally less concern than in extreme cases in the context of OSDC[10]. Thus, while we are now able to make informed 'reasonable' risk assessment recommendations for tree-based models, the next stage is to expand this work to more complex ML algorithms and use-cases.

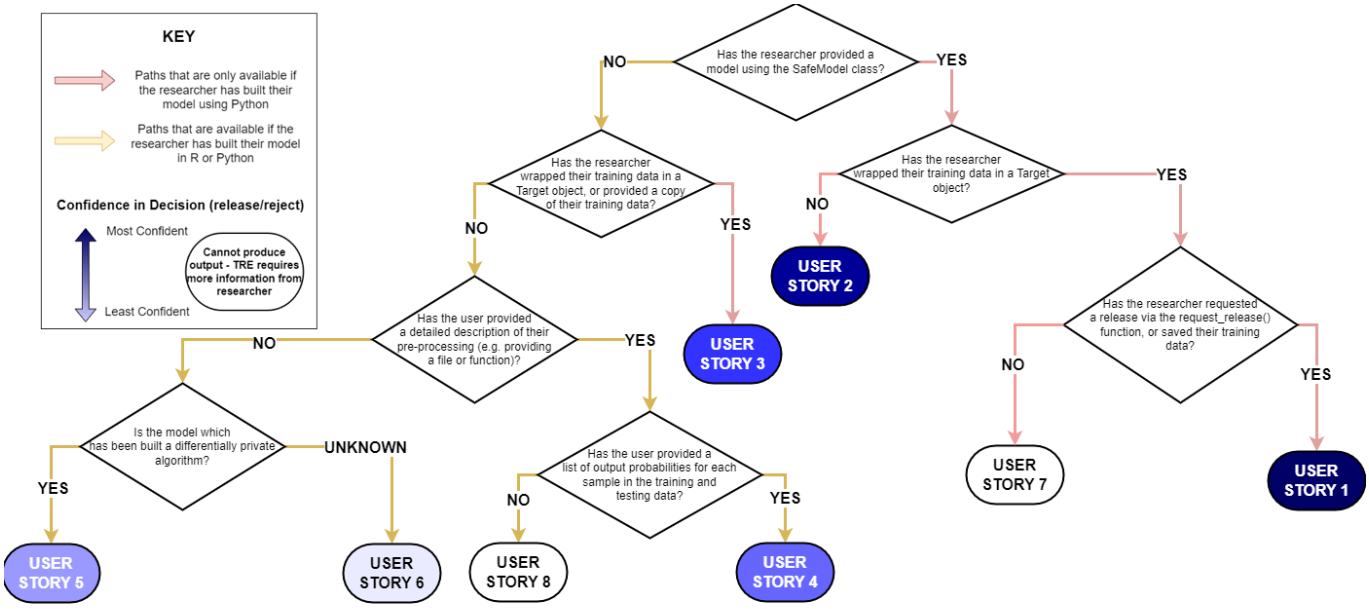


Figure 2: User Stories for automated risk assessment of ML models

2.5. Public Involvement and Engagement

Public engagement activities have been detailed in a separate report. Here we focus on the main outcome of these, which was the creation of a Consensus Statement detailing a set of eight principles surrounding the use of (semi) automated tools for checking the disclosure risk of outputs from confidential data.

These principles were created through engagement with public panels and stakeholders, as well as a review of existing literature on output checking. As of 30/10/2023, the following organisations have agreed to adopt the consensus statement and its principles:

• UWE • OpenSAFELY at the Bennett Institute, University of Oxford • OpenSAFELY Digital Critical Friends Group • Office for National Statistics	• Research Data Scotland • University of Durham • DaSH – Grampian Data Safe Haven • HIC – Health Informatics Centre at University of Dundee	• Public Health Scotland (eDRIS) • ICPSR- Inter-University Consortium for Political and Social Research • EPCC • HDR UK

[10] These concerns were highlighted in the GRAIMatter report available at 10.5281/zenodo.7089490

2.6. External Engagement

The aim of this work package was to engage and ensure SACRO development was shaped by the voice of the TRE practitioners and data services, this would ensure the work stays usable and relevant in a real-world setting. To achieve this a series of workshops and 'hands-on' demonstrations and one to one meetings enabled the direct embedding of SACRO in research organisations- allowing for the tool to be trialled and gain uptake from both the frontline staff and the managers. Full details of all the activities below, and of TRE feedback sessions, are available in the sharepoint site (footnote 1), and will be made publicly available in a suitably anonymised format.

2.6.1. Steering group

A steering group was established at the outset of the project and comprised of senior managers of data services (both nationally and globally) and public representatives.

- Chair: Simon Parker (German Cancer Research Centre DKFZ),
- External steering group members: Bill South (Office for National Statistics), Maggie Levenstein (ICPSR), Richard Welpton (ESRC), Stuart Balint (DARE UK), Fergus McDonald (DAREUK)
- Public Representatives: Nigel Mead, Mary Mancini Fabiola Detari
- SACRO work package leaders: Elizabeth Green (UWE), Amy McNeill (UWE), Jim Smith (UWE), Katie Oldfield (Research Data Scotland), Pete Stokes (OpenSAFELY), Layla Robinson (Research Data Scotland)

This group met a total of 3 times (March, May and August) and provided critical feedback to the SACRO team.

2.6.2. Workshops

As part of the project we held a number of open virtual workshops throughout the project allowing stakeholders to engage with the project and the development of SACRO alongside the SACRO team sharing current findings of the project. The workshops were formed as reverse science cafes- a method which focuses on the participants/ workshop attendees providing input and discussion (thereby informing the future of the project). A total of 7 workshops were held with a wide reach of stakeholders, varying between 10 and 42 attendees (mean 24).

2.6.3. Engagement

Throughout the start and to the delivery of the project we have held a number of one-to-one meetings with TREs, to gain feedback on the development and implementation of SACRO outputs. These meetings included project TREs, and a range of national and international TREs. We also met to share findings and ideas with bodies who are deploying related approaches, specifically Datashield (PI is co-chair of the CoI group), Eurostat (who are using the 'Stata-acro' prototype and have embedded it in their new remote access system) and at the Bundesbank and Statistics Canada who have their own bespoke language-specific tools (but not an equivalent of the viewer).

These targeted meetings were accompanied to presentations to wider meetings such as HDRScotland, TRE-UK conference, and SCADR monthly meetings, and at the UNECE Expert Meeting on Statistical Disclosure Control. Those papers on the theory and toolkits (footnotes 3 and 6 respectively) raised support for the CoI with interest expressed by Eurostat, Central banks (Italy, Spain, Germany), and Stats Netherlands. Eurostat (who supported the initial proof-of-concept) are planning to deploy SACRO and now include coverage of SACRO in their standard 'ESTP' training on output checking.

Future Plans: This external engagement work was instrumental in the formation of the CoI proposal and will be continued through that forum. One early activity will be to establish mechanisms for recording and monitoring the uptake/impact of the SACRO outputs.

3. Impact

Impact of the work done to create the new conceptual framework for OSDC manual is substantial and will affect the implementation of the '5-Safes' even in TREs where the toolkits are not deployed. Intended to be foundational (authoritative, evidenced, structured), the 'statbarns' model is a radical new approach to SDC. It has been trialled in the user and output checker courses, and presented to peers at conferences, with positive feedback. The review of existing literature prompted by the development of the guide has changed a number of core guidelines; coupling this activity with implementation and testing also uncovered a number of areas where OSDC was incorrect or not sufficiently defined.

This conceptual work is already feeding into the OSDC landscape. The SDAP group plans to rewrite their popular manual based on the new core. The new model is already being incorporated into changes in UK national training and accreditation courses for researchers and output checkers to include SACRO materials. These courses affect around 2000 people annually in the UK, plus more internationally. As of October 2023, Eurostat's training courses for staff and researchers now cover the SACRO toolkits.

We also expect this to be a sea change in the way output checking is viewed. As we have argued in research papers, output checking should be seen as an operational process, not a statistical one. SACRO reinforces that by moving the day-to-day, low risk work best done by computers into a computer system, so freeing up specialist resources to concentrate on the statistical queries when they arise. This is expected to lead to changes in the way TREs may be staffed in future. ESRC's Future Data Services strategic review is closely observing how SACRO is impacting on data services (not just TREs), and its recommendations will be looking to support initiatives such as SACRO. Other specific examples include a workshop being held in Feb 2024 with confirmed attendance from all major UK 'admin' TREs (ONS, UKDS, RDS, and IDS) to review their output processes. A further international workshop is to be hosted by Bundesbank in Spring 2024 to explore implications of SACRO theory and tool itself.

The consensus statement provides an overview of the project and outcomes in a relatively accessible way and this has been explicitly supported by a number of leading organisations. One key outcome will be the number of organisations adopting the statement and principles. As the variety of users of TREs increases to include more commercial organisations, there is clearly an obvious benefit from providing a set of clear consistent principles and guidelines, with public backing, that underpin how TREs will apply output checking wherever there is doubt (especially around AI). This becomes especially useful in the case of federated analytics, where TREs Governance approval may benefit from agreed ways of working. This is also something we will continue to build on through the Community of Interest Group.

The likely impact of the SACRO software being supported in TREs and used by researchers is substantial. Judging by the interest shown whenever we have demonstrated SACRO, and the increasing levels of support for the Community of Interest, it seems reasonable to assess that before the end of 2025 SACRO software will be available in the majority of UK TREs, and many internationally. While we are confident that SACRO has achieved its main technical aims, achieving impact now rests on addressing a range of blocks to adoption by TREs – and most importantly researchers, and on continued maintenance and improvement of the code repositories.

DARE UK has awarded seed funding for a Community of Interest group, whose remit may be broadly stated as doing the 'soft' and organisational work needed to achieve, **and put in place measures for quantifying**, impact from the toolkit. Key metrics are likely to include:

- The numbers of researchers using SACRO, and their feedback on whether the instant feedback it provides was positive and helped them reduce the number of output requests that were rejected by TREs.
- The volume of requests processed by TREs within the SACRO viewer, and their feedback on whether it helped provide a better, more timely service with improved transparency and auditing. While this is not currently an issue for some TREs, at the time of writing output checking can take over ten days at some busier TREs, and is typically longer for ML outputs.

- The number of TREs who feel able to support the training of ML models for egress, through a combination of SACRO's aisdc risk assessment tools, incorporating the messages from the 'user stories' from project inception, and the possibility of a 'pool of expertise' to assist in making judgements about complex cases. We have already provided support and advice to 3 of the 5 TRE partners within SACRO. Discussions with TREs and HDR UK suggest it is not unreasonable to see the number of TREs who have been supported reaching the tens within 2-5 years.

The example of considering what is needed to risk-assess a trained ML model from the project outset also holds for federated analytics, and we would argue needs to be a key part of ongoing DARE UK initiatives. For example, certain types of queries (especially those producing 'position' statistics such as medians, or figures) can only be reliably performed and risk assessed on pooled data. Also, there is a general move within TREs away from highly restrictive 'rules-based checking' (as systems such as DataShield currently support), towards principles-based OSDC which can facilitate the publication of a wider, flexible range of outputs more suited to modern research needs that are safe despite breaking restrictive rules.

Therefore, achieving consistency, and above all enabling research for societal good, requires a closer investigation of what it means ***in practice*** to embed SACRO-like functionality within a federated architecture.

4. Relationship to Federated Architecture Blueprint

From the outset all work in this project has been designed to fit within DARE UK's vision for a federated architecture blueprint. For example, all important data is held in appropriate structured machine readable format such as .yaml (for the TRE risk appetite, and ML user story specifications, which needs to be easily human-editable) or JSON (all the information needed to perform a risk assessment of, and then release, results from an analytic query, and the results of running different attacks on Machine Learning models).

This means that SACRO could accommodate the different types of conceptual data spaces in the blueprint. In data pooling mode (e.g., the `Teleport' model) it can sit on the virtual TRE accessed by the researcher. Alternatively, for federated analytics, (e.g., the `TRE-FX' model) it could be split so that acro-queries are distributed to different data holders, where they are handled by acro instances. These would return the analytic results, OSDC analysis, and each TRE's risk appetite to a "SACRO-aggregator" on the researcher's 'host' TRE. There they can be viewed by the researchers and the TRE output checker.

These two extremes emphasise the value of consistent approaches to thinking about and defining risk appetite that support automated reasoning within a principles-based approach to 'Safe Outputs'. There are some gaps that need exploring around query distribution and aggregation, and how these would be accepted by TREs. The benefit is that SACRO lets people focus on processes for how they are going to guarantee they have safe outputs, so helps a focus on how the distributed analytics need to work right from project inception.

As far as we are aware, there has not been much public engagement around the linking of data in the context of a federated approach. It would be interesting and valuable to explore this with groups of people to understand how members of the public feel about their data being used in this way and any concerns there might be and inform the approach. We would be well placed, following the work undertaken through the SACRO project, to be involved in this and next steps around this, such as an updated or broader consensus statement.

Finally, we note that proposals such as TRE-FX has a 'data use register' in their transparency layer. If this were sufficiently detailed, and could be combined with the development of a suitably rich standard for dataset metadata that permitted automated reasoning, this could facilitate the 'holy grail' of automated checking for 'differencing risk'- i.e. that a disclosive combination could be created from two individually safe outputs could be disclosive– a major headache for all TREs with no current satisfactory solution.

5. Acknowledgements

This work is funded by UK research and Innovation, [Grant Number MC_PC_23006], as part of Phase 1 of the DARE UK (Data and Analytics Research Environments UK) programme, delivered in partnership with Health Data Research UK (HDR UK) and Administrative Data Research UK (ADR UK)

This document may be accessed via DOI: 10.5281/zenodo.10055365.

The following project members who contributed to this report

Name	Affiliation	Orcid id (where available)
J. E. Smith	University of the West of England	0000-0001-7908-185
M. Albashir	University of the West of England	-
S. Bacon	Bennett Institute, University of Oxford	0000-0002-6354-3454
B. Butler-Cole	Bennett Institute, University of Oxford	0000-0002-8890-2767
J. Caldwell	Electronic Data Research & Innovation Service (eDRIS), Public Health Scotland	0000-0003-2326-1599
C. Cole	University of Dundee	0000-0002-2560-2484
A. Crespi Boixander	University of Dundee	0000-0002-1576-9723
E. Green	University of the West of England	0000-0002-5199-9534
E. Jefferson	Health Data Research, UK	0000-0003-2992-7582
Y. Jones	NHS Scotland	0000-0003-0169-6364
S. Krueger	University of Dundee	0000-0002-5219-1959
J. Liley	University of Durham	0000-0002-0049-8238
A. McNeill	University of the West of England	-
K. O'Sullivan	University of Aberdeen	0000-0002-9225-4942
K. Oldfield	Research Data Scotland	0009-0000-2941-5662
R. Preen	University of the West of England	0000-0003-3351-8132
F. Ritchie	University of the West of England	0000-0003-4097-402
L. Robinson	Research Data Scotland	0009-0006-2381-3428
S. Rogers	NHS Scotland	0000-0003-3578-4477
P. Stokes	Bennett Institute, University of Oxford	0000-0002-2486-8969
A. Tilbrook	University of Edinburgh	0000-0002-0294-2101
P. White	University of the West of England	0000-0002-7503-9896