DARE UK (Data and Analytics Research Environments UK)

Statistical Disclosure Control- Reducing Barriers to Outputs from TREs: (SDC-Reboot)

The WHY

Introduction

The DARE Driver project Semi-Automated Checking of Research outputs (SACRO), and before it, the DARE Sprint exemplar Guidelines and Resources for AI Model Access from Trusted Research Environments (GRAIMatter), delivered a suite of guidelines and tools for Output Statistical Disclosure Control (OSDC). These have established considerable interest and appetite within the UK and international community, which we now need to build on in a sustainable way.

It is now also appropriate and desirable to provide evaluations of alternative paradigms and technologies, that can inform TREs choosing between, for example, the principles-based manual approach supported by SACRO, the fully automated, strict rules-based approach implemented within DataSHIELD , and various toolkits developed in the Machine Learning (ML) field which are not intended for OSDC but may have some use.

Recognising the significance of the topics, DARE identified '*Technology evaluation and comparisons in … Automated output checking for TREs (including evaluation of AI model outputs)*' as one of its priority areas for community groups.

This proposal is timely given:

- The level of interest in the SACRO tools arising from presentations at HDR, DARE, UK TRE Community and other national and international meetings
- The rising numbers and urgency of 'referrals' from TREs who have ML models that they need to risk-assess but do not feel equipped to do.
- The overall DARE vision of interconnected federated analytics wherein the need for disclosure control will escalate significantly and hence need to be supported by thoroughly tested and 'community-approved' automated tools for OSDC.

The challenges this group will address are:

1. Establishing a user community to:
    a. continue ongoing maintenance and development of the tools arising from the SACRO project
    b. embrace and evaluate other tools, both extant and as they may arise; and
    c. establish protocols and mechanisms for evaluating different technologies related to OSDC.
2. Identifying and addressing blocks to impact- such as community created and delivered training for both researchers and TREs, and assurance for IT Security manager and governance groups at TREs.

3. Enabling TREs to support the creation and egress of machine learning models trained on confidential data. This long-term challenge has multiple facets which need to be addressed simultaneously:
    a. on-going development and maintenance of risk assessment/minimisation tools (for researchers and TREs)
    b. creating the *right* training resources for both TREs and researchers
    c. addressing the skills-gap at TREs around ML model privacy risks
    d. establishing a pool of experts and means of engaging them, that TREs can call on for help in assessing outputs they don't feel confident doing 'in-house' particularly the more complex ML cases.
4. Ensuring that the OSDC community, and any resources it develops, are informed by other community interest groups and developments and will meet the future demands of federation and distributed data governance and OSDC.

The WHAT

Objectives

This community group primarily addresses the DARE theme "Capability and Capacity"; however, the work involved necessarily also impacts on and will be informed by other work on "Demonstrating Trustworthiness", "Data and Discovery", and 'Core Federation Services".

It differs from other groups in the area through its focus on 'Safe Outputs', since without the capability to provide this assurance a key part of the 'Five-Safes' disappears, and DARE has recognised that manual checking currently presents a bottleneck. Moreover, studies before and during the DARE 'GRAIMatter' project established that many (possibly all) TREs do not feel equipped to assess the disclosure risk of machine learning models trained on sensitive data and will need expertise they can draw on to provide support when they lack the technical understanding in-house.

The community group will initially work on four parallel threads, all underpinned by the Terms of Reference and the establishment of governance mechanisms for the open-source repositories. Naturally, there will be an overlap between involvement in these threads, and in the longer term, we expect that new foci may emerge and others diminish in priority.

Note that although we use SACRO for brevity, this should be taken to include a range of other tools for (semi) automated OSDC, both extant (e.g. DataSHIELD, ML-Privacy-Meter) and future developments.

Focus 1: Conceptual development and guidelines:

- Rationale: SACRO substantially changed perspective on Output Statistical Disclosure Control (OSDC) through the creation of a framework and taxonomy, and we now need to revise other materials and get community agreement. We also need to make these available in a form that they can be adopted by other tools so as (i) to allow principled comparison between technologies and (ii) promote consensus around best practices in recognising and mitigating risks
- Activities: (i) Workshop to: review and adopt material; identify the need for further material e.g., alignment with SDAP manual (currently under revision); identify areas and priorities for future development and collaborations.

- **Outcomes**: Revised perspective on OSDC; established expert group to take forward future changes & roadmap

Focus 2: Removing barriers to adoption by researchers:

- **Rationale**: SACRO's testing has confirmed that the principle works. However, natural adoption by users is currently very limited (although Eurostat report a growing uptake of the predecessor *acro* Stata tool). Feedback from the community is *where possible* to encourage, rather than force researchers to use specific toolkits (as per DataSHIELD and tools at Stats Canada). Therefore, what is needed is community-led development of resources to enable and motivate researchers, such as training, videos, example code, and enhanced help documentation, that can be accessed both *prior* to a 'TRE research session' and *during* the session (when external access is more limited)

- **Activities:** (i) Series of online and in-person workshops to design and approve resources to be developed, such as user-centred training materials. (ii) Establishment of 'community researcher mentors' who can provide on-going support, e.g., through regular advertised 'drop-in help sessions' and a central email support service.

- **Outcomes:** Roadmap for user adoption; resources; sustainable on-going researcher support network.

Focus 3: Enabling adoption

- **Rationale:** The TREs involved in the SACRO project all tested SACRO but achieved different levels of deployment within their secure environments, mainly due to separate infrastructure changes. Additionally, 3rd party tools like DataSHIELD have also expressed an interest to adopt the framework and taxonomy underpinning SACRO in an effort to harmonise best practice across semi- and automated tools for OSDC. We now need a sustainable network of support for TREs and SDEs in: making a case for deploying automated checking (e.g., around IT governance/ software risk analysis); agreeing methodologies and tools to support evaluation (e.g., tools for 'reverse engineering' previous researcher code into the 'SACRO' framework to permit side-by-side comparison); on-going development of TRE-facing training materials; (iv) adding capability.

- **Activities:** (i) Monthly online/in-person/hybrid workshops to design and approve resources such as installation guides, governance protocols for code repositories, and mechanisms for community prioritisation of 'wish-lists'. (ii) Weekly online 'drop-in' help sessions for TREs at different stages of deployment and evaluation. (iii) Identification of `champions' among TRE community members who can provide peer-mentoring and mechanisms for building this out sustainably. (iv) Outreach to other (DARE) community groups, especially regarding federated analytics and Information Governance.

- **Outcomes:** White paper comparing different technologies for automated OSDC. Governance structures for repositories. Range of materials and mentoring support for new organisations. Ongoing support and development of SACRO code repositories.

Focus 4: Risk assessment of Machine Learning models

- **Rationale**: SACRO and, prior to that, GRAIMatter have established a range of guidelines and mechanisms for automatically assessing the disclosure risk of trained ML models according to several different metrics. However, this is a rapidly moving field, and conceptual gaps still exist between the ways that 'traditional OSDC and ML-privacy research consider risk, which SACRO has only partially been able to address. We need to establish of a community of expertise in interpreting ML risk metrics and explaining them to researchers and governance teams. This goes some way to addressing the skills gap identified during GRAIMatter – that it is probably not realistic to expect all TREs to maintain this expertise 'in-house'.

We also recognise the vital role of public confidence in the safe and appropriate use of their data 'for AI', and ensure that we are aware of, and take into consideration, PIE activity in this area.

- Activities: (i) Establishing a sustainable series of workshops around 'ML privacy risk in the context of TREs'. (ii) Establishment of a 'Community of Expertise' – a pool of experienced people and archive of experience around assessing specific ML models. (iii) Ongoing development and support for the AI-SDC code toolkit for ML risk assessment - to include new forms of attack as the field develops.
- Outcomes: Sustainable community of expertise enabling: closer alignment of OSDC and developments in risk assessment from ML researchers; ongoing support and extension of risk-analysis toolsets; mechanisms for providing practical advice and support to researchers and TREs.

Outcomes

The initial foci and their proposed outcomes are listed above and may be summarised as:

- Alignment of conceptual framework and taxonomy of outputs with current and future training resources developed elsewhere.
- Establishing sustainable processes for the community-led design and implementation of resources to address methodological, 'practical' and training needs for the evaluation, deployment and increasing uptake of automated methods for OSDC.
- Establishing a 'Community of Expertise around the OSDC of Machine Learning models, including training and tool resources for TREs and researchers, a living archive of best practice, and a pool of people who can be drawn on to provide decision support for TREs around ML models.

The HOW

Participation/Collaboration

The community will initially include people from the following groups:

- Project leaders from a range of related projects such as TREvolution, GRAIMAtter, SACRO and DataSHIELD, and community working groups around AI.
- TREs, SDEs and TRE-hosting organisations such as SAIL keen to evaluate (semi) automated tools for OSDC for 'traditional' outputs and deploy them to address capacity and allow their staff to focus on more challenging cases.
- TREs keen to establish working mechanisms allowing them to support the creation and release of ML models trained on the data they hold.
- Computer scientists interested in the development of such tools and resources both for individual TREs and to support federated analytics.
- AI/ML researchers interested in theoretical concepts and algorithms for assessing and quantifying the privacy disclosure risk of trained models *in the context of TREs*.
- Researchers, practitioners, and training-providers involved in the wider development of the concept of disclosure risk and what constitutes 'Safe Outputs'.

- Public engagement practitioners focussing on public perceptions of trustworthiness and the ‘balancing the risk’ between privacy leakage and potential public good of research findings and risk of not supporting such research.

As the community becomes established and increases its outreach, we expect to widen the representation to include a range of organisations ranging from charities to National Statistics Institutes.

Mechanism

The group will meet monthly in a mixture of face-to-face, hybrid and online meetings. Each focus stream will self-organise under the lead of a co-chair and may meet more often. They will organise targeted meetings with a wider audience, including annual workshops co-located with other projects/organisations’ initiatives to maximise attendance. The Chair and co-chairs will meet fortnightly to review progress and balance effort between focus streams. As this is intended to be a living community, any member may propose new activities of focus or changes to existing foci for discussion at the monthly meetings.

Potential members

Representatives from the following organisations have attended in-person SDC-Reboot events and/or signed up to the SDC-REBOOT email list.

SDC Reboot Co-Chairs

First Name	Surname	Co-Chair	Email	Organisation
Jackie	Caldwell	Co-Chair	Jackie.Caldwell3@PHS.SCOT	Public Health Scotland
Lizzie	Green	Co-Chair	elizabeth7.green@UWE.AC.UK	UWE
Lewis	Hotchkiss	Co-Chair	lewis.hotchkiss@chi.swan.ac.uk	Dementias Platforms Uk
Katherine	O'Sullivan	Co-Chair	k.k.osullivan@SHEFFIELD.AC.UK	Sheffield University
Felix	Ritchie	Co-Chair	Felix.ritchie@UWE.AC.UK	UWE
Simon	Rogers	Co-Chair	simon.rogers@NHS.SCOT	NHS Scotland
Jim	Smith	Co-Chair	james.smith@UWE.AC.UK	UWE
Becca	Wilson	Co-Chair	becca.wilson@LIVERPOOL.AC.UK	Liverpool University/DataShield

Organisations
Aberdeen University
Arjuna Technologies Limited
BioResource - NIHR
Cancer Research UK
CPRD (Clinical Practice Research Datalink)

Crick Institute
DARE UK
Data loch /University of Edinburgh
Data Shield/ University of Liverpool
Dementias Platforms Uk
Edinburgh University
Eurostat
Genomics England
GESIS
Health and Social Care Northern Ireland
Health Data Research UK
Health Informatics Centre, Dundee
Health innovations East
HMRC
Lancashire Teaching Hospitals NHS Foundation Trust
Liverpool University/DataShield
Manchester University
Medicines and Healthcare products Regulatory Agency
Mondragon University, Bilbao
MRC Epidemiology Unit, Cambridge University
NHS Scotland
NHS England
NHS South,Central and West Commissioning Support Unit
NISRA
Office for National Statistics (ONS)
PHPS, University of Liverpool
Public Health Scotland
SAIL Databank
Sheffield University

Southampton University
Swansea University
UCL Hospitals
University College London
University of the West of England

** Note, please do not hesitate to point out gaps in the current DARE UK set of strategic themes and/or recommendations that the programme should consider as it continues to evolve these. Community feedback and input is welcomed.*