

University of Dundee

A review of trusted research environments to support next generation capabilities based on interview analysis

Kavianpour, Sanaz; Sutherland, James; Mansouri-Benssassi, Esma; Coull, Natalie ; Jefferson, Emily

Publication date:
2021

Licence:
CC BY

Document Version
Early version, also known as pre-print

[Link to publication in Discovery Research Portal](#)

Citation for published version (APA):

Kavianpour, S., Sutherland, J., Mansouri-Benssassi, E., Coull, N., & Jefferson, E. (2021). *A review of trusted research environments to support next generation capabilities based on interview analysis*. JMIR Publications Inc.

General rights

Copyright and moral rights for the publications made accessible in Discovery Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from Discovery Research Portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain.
- You may freely distribute the URL identifying the publication in the public portal.

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

A Review of Trusted Research Environments to Support Next Generation Capabilities based on Interview Analysis

Sanaz Kavianpour, James Sutherland, Esma Mansouri-Benssassi, Natalie Coull,
Emily Jefferson

Submitted to: Journal of Medical Internet Research
on: September 20, 2021

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 4



A Review of Trusted Research Environments to Support Next Generation Capabilities based on Interview Analysis

Sanaz Kavianpour¹ PhD; James Sutherland² PhD; Esma Mansouri-Benssassi² PhD; Natalie Coull¹ PhD; Emily Jefferson² PhD

¹Abertay University Dundee GB

²Health Informatics Centre University of Dundee Dundee GB

Corresponding Author:

James Sutherland PhD
Health Informatics Centre
University of Dundee
Dundee
Dundee
GB

Abstract

Background: A Trusted Research Environment (also known as a Safe Haven) is an environment supported by trained staff and agreed processes (principles and standards) providing access to data for research whilst protecting patient confidentiality. Accessing sensitive data without compromising the privacy and security of the data is a complex process.

Objective: This paper presents the security measures, administrative procedures and technical approaches adopted by TREs.

Methods: We contacted TRE operators, 20 of whom, in the UK and internationally, agreed to be interviewed remotely under a non-disclosure agreement and to complete a questionnaire about their TRE.

Results: We observed many similar processes and standards which TREs follow to adhere to the Seven Safes principles. The security processes and TRE capabilities for supporting observational studies using classical statistical methods were mature and the requirements well understood. However, we identified limitations in the security measures and capabilities of TREs to support “next-generation” requirements such as wider ranges of data types, the ability to develop artificial intelligence algorithms and software within the environment, the handling of big data, and timely import and export of data.

Conclusions: We found a lack of software/automation tools to support the community and limited knowledge of how to meet next-generation requirements from the research community. Disclosure control for exporting artificial intelligence (AI) algorithms and software was found to be particularly challenging where there is a clear need for additional controls to support this capability within TREs.

(JMIR Preprints 20/09/2021:33720)

DOI: <https://doi.org/10.2196/preprints.33720>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

Please make my preprint PDF available to anyone at any time (recommended).

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in [http](#)

Original Manuscript



A Review of Trusted Research Environments to Support Next Generation Capabilities based on Interview Analysis

Sanaz Kavianpour, James Sutherland, Esma Mansouri-Benssassi, Natalie Coull, Emily Jefferson

Abstract

Background: A Trusted Research Environment (also known as a Safe Haven) is an environment supported by trained staff and agreed processes (principles and standards) providing access to data for research whilst protecting patient confidentiality. Accessing sensitive data without compromising the privacy and security of the data is a complex process.

Objectives: This paper presents the security measures, administrative procedures and technical approaches adopted by TREs.

Methods: We contacted TRE operators, 20 of whom, in the UK and internationally, agreed to be interviewed remotely under a non-disclosure agreement and to complete a questionnaire about their TRE.

Results: We observed many similar processes and standards which TREs follow to adhere to the Seven Safes principles. The security processes and TRE capabilities for supporting observational studies using classical statistical methods were mature and the requirements well understood. However, we identified limitations in the security measures and capabilities of TREs to support “next-generation” requirements such as wider ranges of data types, the ability to develop artificial intelligence algorithms and software within the environment, the handling of big data, and timely import and export of data.

Conclusions: We found a lack of software/automation tools to support the community and limited knowledge of how to meet next-generation requirements from the research community. Disclosure control for exporting artificial intelligence (AI) algorithms and software was found to be particularly challenging where there is a clear need for additional controls to support this capability within TREs.

Keywords: Data Safe Haven, Health Data Analysis, Trusted Research Environment (TRE)

1 Introduction

A Trusted Research Environment (TRE), also known as a Data Safe Haven, is a secure environment designed for approved, and named researchers to access sensitive data, where access to specific datasets is provided to approved research projects. To protect the confidentiality and privacy of the data, TRE providers and researchers using the environments generally follow a set of TRE principles. Such principles have developed over time, for example the Scottish Safe Haven Charter and the Health Data Research Alliance Trusted Research Environment Green Paper [1-3].

The objective of a TRE is to provide safe and trustworthy access to data for research. Controls are generally applied to both the import and export of data to protect the privacy of subjects and the integrity of the environment itself. Within the secure environment, researchers can analyse this data using a set of advanced analytics tools, for example R and SPSS. Some TREs also offer the researcher the capability to program within the environment and support the development of new Artificial Intelligence (AI) and to apply Natural Language Processing (NLP) for the analysis of unstructured text.

Many of the first TREs were developed to host health data. For example, in the United Kingdom (UK) there are several TREs that were established to host health records from the National Health Service (NHS), the publicly funded healthcare system of the UK [4-8]. A similar model has since been adopted to provide secure access to many other non-health datasets [9-11]. Many TREs now regularly hosts both health and non-health data.

Providing researchers with access to sensitive data sources without compromising the privacy and security of the data is a complex process. Historically TREs have mainly supported observational studies on text-based structured data using standard statistical packages. There is a growing requirement from the research community (academic and industrial) for TREs to provide additional capability beyond simple support for observational data statistical analysis, but without compromising the security or privacy of the data. Here we term these requirements Next Generation TRE capabilities. These include:

- Support for big, non-structured data (such as genomic and imaging data which can be several terabytes in size)
- Ability to parallelise computational jobs to either a High-Performance Computing (HPC) cluster or GPU farm
- Support for software development within the TRE
- Freedom to install software packages of researcher choice
- Ability to export software and artificial intelligence (AI) algorithms from the environment
- Ability to connect to certain internet locations, e.g. code repositories (GitHub)

This research aims to understand the state-of-the-art in supporting next generation TRE capabilities, understand the existing technical security measures that have been adopted and how widely, and understand the limitations in existing controls and processes, where active research is required to develop novel methods.

The findings are based on interviews conducted with fifteen TRE providers in the UK and five TRE providers across Canada, Australia, and Europe. Each interview took approximately two hours, using a set of questions that were designed to cover TRE controls and next generation capabilities (see Appendix A).

Building upon the five safes model [12], the recent HDR UK Green Paper [2] describes 7 “Safes”: Safe people, Safe projects, Safe setting, Safe computing (an extension of Safe setting), Safe data, Safe outputs, and Safe return (extending the TRE definition). This research focuses on a subset of the controls which would support next generation TRE capabilities (data, outputs, settings, computing, and people). Table 1 provides a summary of the topics that were discussed with the participants during the interview, under each of our subset of the 7 Safes.

The interview participants were recruited mainly from the technical TRE infrastructure teams. A Non-Disclosure Agreement (NDA) was signed by the project parties, to provide assurance for any participants who chose to disclose information of a confidential or proprietary nature. We have anonymised the responses in this paper and grouped the responses under the different Safes. Where relevant, we also present our analysis of the TRE limitations identified, and recommendations that we believe could help. In Section 7, the participants' recommendations for TREs improvement are explained. The paper concludes the features that are significant to improving next generation TRE capabilities in section 8.

Table 1. Safes and Discussion Points

Safes	Discussion Points
Safe Data	<ul style="list-style-type: none"> • Tools and techniques used to manage and reduce the potential risk of re-identification by applying disclosure control to all data imported to the TRE

Safe Outputs	<ul style="list-style-type: none"> • Types of data that can be exported from the TRE. • Future plans to enable export of additional types of data. • The process used for checking disclosure control on data to be exported, including frequency, restrictions. • Software, and any manual checks used for disclosure control.
Safe Settings	<ul style="list-style-type: none"> • Standard build of the TRE (including computing power, operating system, and software). • Maximum computing power offered to TRE “power” users. • Security measures employed to mitigate the risk of unauthorised access, data loss, and misuse. • Rules regarding the importing of data or code (including libraries) into the environment. • Support for federated queries of data from external sources.
Safe Computing	<ul style="list-style-type: none"> • Use of private or public cloud.
Safe People	The controls put on the people who use the TRE, and whether access to the environment must be via a recognised “trusted” organisation.

2 Safe Data

The principles of Safe Data relate to the data allowed to be imported to the TRE. Good practice indicates that such data should be of high quality and pseudonymised/anonymised [2]. Researchers accessing the TRE should only be able to access data necessary for their research project: the work of protecting data begins by applying Disclosure Control to control and assess data provided to the researchers within the TRE, before a second pass at the export stage regulates the data those researchers are allowed to disclose publicly within their research output.

Many TREs provide a service to link and anonymise data from different data sources. Such research projects often require data governance approval before researchers can access the data. TREs often require researchers to sign Data User Agreements, which reinforce the rules and consequences for violations. During the interviews, participants were asked to discuss existing tools used to support Safe Data and discuss views on future solutions. These are discussed below.

3.1 Breaches or Near-Miss Incidents

Each TRE participating has processes in place for handling data breaches, including reporting them to the appropriate authorities when required. None reported any actual reportable data breaches. Two of the participants acknowledged that there could be breaches of procedures (these are different to data breaches and do not need to be reported to the ICO), where a researcher may request to export data that is not permitted, perhaps due to researcher error. In these cases, the procedural breach would be addressed with a process for review, formal warning and retraining if necessary. One TRE described an incident in which a field that normally contains facility/hospital names was provided to a researcher, and some records included private addresses in this field. This data was then retracted, and the data resupplied with a placeholder for private addresses. In this example the data was not released outside of the TRE, but researchers did see the potentially identifiable data within the TRE environment.

2.2 Tools/Techniques Used to Manage the Risks of Re-identification

Most participants do not rely on special purpose disclosure control tools, but on their analysts' knowledge and communication about the purpose of the project, and nature of the data. Ultimately, TREs exist with the aim of providing "Safe Data" sufficient for the project's needs. Generally, patient IDs are replaced with either safe-haven or project specific identifiers; where possible, other identifying information is redacted or reduced in resolution, for example replacing a date of birth with a year or month of birth. The combination of data across projects is not allowed, and made impractical through the use of project-specific pseudonyms in place of original identifiers.

Four participants explicitly measure disclosure risks at the import stage for each project, though all participants acknowledged that disclosure risk is an important factor when evaluating a new project application. As part of this process, one TRE has all data scrutinised by an external organisation, using the k-value (the minimum number of individuals sharing any combination of identifying characteristics) to quantify the re-identification risk for each dataset [13]. For example, a dataset classifying patients by age and gender with a minimum of 4 patients in each category has $k=4$. Grouping patients into larger age bands would raise the k-value, reducing risk at the expense of reduced data resolution. TREs apply a similar measure as part of the export process, usually requiring a minimum of 5-10 individuals in any output grouping to ensure no individual can be re-identified, as detailed in section 3.2.

Four participants reported using tools to support checks in the import and export process. Each used a different tool:

- SDCMicro, (as discussed in Section 2.6) is a free R-based open-source package which assesses the risk of a data set containing identifiable data using various risk estimation methods [14].
- Privitar's Data Privacy Platform product applies user-defined policies for filtering and transforming data, including adding random 'noise' to numerical values in configurable ways to reduce identifiability [15].
- The Custodix Anonymisation Services (CATS) de-identification platform [16] provides both assessment and de-identifying transformation of various types of data (e.g., CSV, DICOM).
- Privacy Analytics Risk Assessment Tool (PARAT) [17] is used by one TRE to risk-assess, redact and de-identify source data and also to assess export requests.

While some TREs use existing tools, the costs involved with proprietary tools were a deterrent for many TREs, which did not feel that the cost could be covered by research funds.

Recommendations: There is a need for affordable tools that can be used in TREs to support de-identification of data and assess the risk of re-identification; in particular, building on existing use of the free SDCmicro toolkit should be considered. Data identifiability is not binary [18], and data can be identified indirectly by combining attributes, known as a triangulation attack [19].

2.3 Watermarking Used to Preserve Health Data Privacy

Digital watermarking [20] is embedding subtle identification information into data, sometimes used to trace unauthorised copying of multimedia data, databases, maps and text files. This was discussed with the participants: none use it at present, though two felt it could potentially be a useful control, another stated that it could be useful if the data is shared without the existing controls.

Limitations: Since watermarking entails modifications to the data, albeit subtle, there will be a

risk of causing problems in any analysis done on such modified data. An intentional spelling error, colour change or movement of a landmark may be harmless in cartographic data, but this is more difficult with medical data: a slight shadow on an X-ray could be taken as a significant symptom, an altered dosage from 25mg to 26mg might be an obvious anomaly.

Recommendations: Although watermarking technology does not help to preserve privacy or secure a TRE, it may be valuable in tracing breaches should they occur. This should be further explored.

3 Safe Outputs

Participants were asked which data types can be exported from their TRE and what controls are placed on the export of data. These checks are more extensive than for imported data, adding checks for deliberate attempts to hide data e.g., no white text in the document, or embedded in Stata code, as well as generic checks on the actual data going further than the input checks: usually an absolute prohibition on data regarding any individual (“row-level”).

Generally, the researcher explicitly requests the export of specific files, and these are then reviewed by TRE staff, and in some cases other relevant external parties, for example data owners, prior to permission to export being granted. All participants have instructions that document the manual checks required. The team determines the extent of checks required at a data or project level based on the sensitivity of the data. For example, openly available, public data sets would not typically require an independent referee, but clinical data may require more consideration. For highly sensitive data, more than one member of the review team may be required to check each file.

The checks needed varied between TREs and sometimes between projects, depending on the nature of the data, including how sensitive it is and whether it was consented or not, with release criteria being agreed between data owners and TRE operators, sometimes consulting with research teams for specific situations, then enforced by TRE staff. Many TREs have developed a rule-based framework to categorise projects and data into specific types. For example, open public data could be exported with minimal checks, while for clinical data only aggregate level summary data could be exported. Others used a simpler ‘one size fits all’ approach.

2.1 Export of Individual Level Data

Generally, export of row or individual level data is only permitted for projects where the data is open data already in the public domain, or where specific consent had been provided by the research participants to collect and share the data. In the latter case, respondents indicated that the data controller was most likely to be the PI of the project and they would typically utilise the TRE to securely manage access to their data by the different researchers involved in the project but would choose not to place restrictions on data export. Otherwise, only aggregate statistics can be exported.

3.2 Export of Aggregate Level Statistical Analysis

Most TREs only allow export of aggregate level statistical analysis. For example, in clinical data projects where the Data Controller is an NHS board/trust, researchers are not permitted to export any data relating to specific individuals (even if pseudonymised).

All participants indicated that their TREs allow the export of aggregate level data as graphs or tables, with a minimum number of data points in any table cell or graphical output to reduce

the probability of re-identification of data from small sample sizes (“small cell risk”). Seventeen of the participants have set policies in place: seven of the participating TREs use a minimum of 5 individuals in a cell, with two TREs using 10 and eight more varying the limit depending on the context of the research, and the nature of the underlying data.

It was acknowledged by all the participants that there was a potential risk of re-identification if enough data points are exported from the TRE, known as jigsaw identification or triangulation [21]. While there is a clear need for researchers to export aggregate level statistical analysis, the mitigations employed for these risks vary across different data sets [3]. There are software tools that can be used to estimate the probability of re-identification, discussed in Section 2.2.

Introducing new types of data as export options brings new risks, particularly AI models and software where statistical analysts will be unfamiliar with the nature of such files and manual inspection is ineffective as well as difficult and time-consuming, introducing new security risks. Data could be intentionally concealed within such files, recovered by a third party from an exported model created innocently through means such as membership inference attacks [22] or, with some inside knowledge or collusion with a researcher, inversion attacks to deduce additional information about the training data [23].

In the rest of this section, we discuss the different data types that can be exported from the participants’ TREs, and how this is managed.

3.3 Export of AI Algorithms, Software, and Scripts

Participants were asked about policies regarding the export of software and AI models developed within their TREs. Five do allow export of AI models, eight specifically prohibit this and five more are prepared to consider it in future.

Some participants also plan to support the export of R and Stata scripts in the future, providing they had in place a suitable process for reviewing.

Fourteen TREs permit exporting software source code developed within the TRE. None have been asked to allow compiled executables; two are prepared to consider this with safeguards, most rule this out as too risky. It should also be entirely avoidable, however, by developing the source code outside the TRE then deploying into the TRE for testing [24].

Limitations: Many participants indicated that checking algorithms, software and scripts is very challenging, as a malicious individual can “hide” individual level data within the files. For example, the weights of an AI algorithm are a set of numbers and sensitive data could be embedded in them. This is very difficult to detect, especially if a malicious user disguises the data. It is also possible to include individual level data inadvertently, for example if the AI algorithm is over-trained, and the weights correspond to the data underneath, or if an R-script incorporates the underpinning data. Checking a substantial software project manually is unrealistic.

Recommendations: Developing AI models in TREs without compromising patient privacy requires tools such as those proposed by [25,26] to quantify their risk and vulnerability to attacks (for example Membership Inference attacks [27,28], De-anonymisation attacks [29], Reconstruction attacks [30], Model Extraction attacks [31,32], and Model Inversion attacks [24,33]) and consider integrating privacy mechanisms in the model development to counter these attacks [34]. Best practices guidelines can also help users design robust and safe algorithms, including through auditable and explainable AI [35]. Software tools to check for non-malicious export by comparing individual data within the TRE to the export files is a possibility, but such tools are not currently routinely used by any of the TREs. Barriers to their

usage by TREs include the attack-specific nature of such tools and their high price. For software development (as opposed to AI models, where the training data is an essential input to the end product) exporting the software from the TRE can be avoided entirely by developing outside.

3.4 Automation of Data Export Checks

Although software could theoretically be used to facilitate the data export process, no participant believed that software could currently replace the role of humans for checking export files. One participant questioned whether it would ever be feasible to fully automate all aspects of the process, largely due to concerns about trusting software to do all the necessary checks without human oversight. Many participants felt that the software available is not currently mature enough to manage all the risks, and humans are better at the task, however, they indicated that they would be willing to incorporate software of this nature into the process in the future as the technology evolves.

Two participants reported using automated tools for export checks. One, as noted in section 2.2, used the proprietary PARAT product; another used a simple in-house tool to detect the project-specific identifiers they use, but the main disclosure checks are manual. At present, use of automated tooling seems more prevalent on import than export checks.

Limitations: Manual checks are time consuming and error-prone, with a risk of missing concealed data (steganographic, white-on-white text, undo buffers) as well as causing delays in data release. Although the participants acknowledged that the current DC process could be enhanced with automated tools, there remains significant concerns with relying solely on technology to check export requests, based on the potential ramifications for any unapproved data to accidentally leave the environment and the challenges with checking algorithms. Proprietary tools are expensive and TREs try to keep costs low for academic research.

Recommendations: A hybrid model with automated checks could facilitate and accelerate export and reduce the risk of re-identification, checking more thoroughly for inadvertent and malicious inclusion of data. The tools noted in section 2.2 may be useful in this role as well. Best practice guidance considering methods to reduce the opportunities for malicious data exfiltration could also help. While governance (see section 6) can help to ensure that researchers are trustworthy, malicious attempts to hide data should be considered, e.g., in the event of stolen researcher credentials.

3.5 Frequency of Data Exports

All participants reported that researchers can request data to be exported from the TRE at any time, although the frequency of requests varied significantly per project, with some requiring daily exports and others exporting only at the end of a project.

Participants were keen to explore how the review process of data export can be improved and automated to decrease the review team workload. However, there was concern that more frequent exports increased risks of data leakage. For example, two consecutive releases featuring a subgroup of 26 and 27 patients respectively would each be acceptable in isolation but comparing the two discloses additional information about the additional patient in the second release.

Limitations: As the manual export checking process uses significant staff time, some TREs apply limits on the number of exports, or charge projects more for frequent usage. One participant explained that the volume of export requests allowed is related to the cost model of the

tenancy. For example, one TRE only allowed two releases for MSc/BSc projects.

Recommendations: Due to the different types of data used across the different TREs, and the different types of projects, it is evident there is no “one-size fits all” solution, but rather a solution needs to be flexible enough to facilitate these differences between projects, data sets, and TREs. Automation could help address these resource concerns and increase the speed and frequency that researchers are able to export data. Although human checks are useful, the process has limitations and the risk of human error.

3.6 Potential Gaps in Export Checks

Participants were asked if they perceived any gaps in the export process, and how they thought it could be improved. Fifteen were not aware of any gaps or security concerns. The following concerns were raised by the others:

- Researchers could be creative in finding a way to remove data, for example, using screen capture to exfiltrate data, which would be difficult to detect.
- Manual checks have the potential for human error.
- Due to the variety of data types that would be requested for export, it was difficult to find software that had the functionality to check all file types. This variety also makes it challenging to bring together a review team that has knowledge of where data may be accidentally or deliberately hidden, particularly novel data types. None of the TREs were aware of any existing software tools that could be used for checking algorithmic data export requests.
- Deficiencies in the audit trail make it impossible to see what researchers did in the TRE, as sometimes research may have deviated from the original goal, and this was difficult to detect.

One of the participants mentioned that the manual process could be greatly enhanced by:

- effective training
- ensuring that staff rigorously check outputs
- applying the principles of appropriate frameworks, such as the Seven Safes, and nationally recognised “best practice” (e.g. the Canadian essential requirements for operating data trusts [36]).
- having a collaborative relationship with researchers throughout their project to mitigate and prevent malicious behaviour

4 Safe Settings

The Safe Settings controls cover the infrastructure and associated security measures that should be adopted by TREs. These controls specify that computing power and operating systems should enable a safe setting to sustain both economical scalability of compute for analysis (e.g., images, genomics) and integral data security. Safe Setting controls describe best practices of policies, techniques and security measures and strategies that are required when sharing data for analysis.

4.1 Computing Power and Operating System Offered to a “Standard” User

Generally, TREs take the form of a Virtual Desktop Infrastructure (VDI) – each user gets remote

access to a desktop environment with access to their project's data and appropriate software with which to analyse that data. Most give each user their own Virtual Machine (VM), with fixed resources (particularly memory and processing) isolated from other projects and users, while a few share a multi-user system more directly (known as "session-based" VDI), allowing a user to exploit the full hardware capacity of the host system when needed at the expense of reduced isolation between users and projects.

Most respondents indicated that there were usually standard templates used for the TRE, and the computing power offered for a project would depend on the number of users who needed to access it. One creates custom configurations for every project, two have no flexibility available, and twelve respondents reported that their TREs could scale up depending on researcher requirements. Heavy compute would have higher costs associated with it, which could be a barrier for many research projects. One of the participants mentioned that the maximum computing power configurations depend upon each individual project's budget constraints. Table 2 indicates the different computer power available across the TREs. (Note: some of the participants were not able to answer this question; some use the public cloud, so resources are effectively limited only by budget).

Table 2. Available Computing Power

Processing power	RAM	Storage space	Allocation
1 CPU	8 GB	5 TB	VM power
2 CPUs	8GB	250GB (fast scratch)	VM power
4 CPUs	16GB	~	VM power
1.5 cores	18GB	1TB	VM power
4 to 8 cores	32 to 64GB	60 to 80GB	VM power
4 to 64 cores	~	8GB to 2TB	Host power
16 cores	~	96GB	Host power
dual Xeon processor	~	120GB	VM power
	~2000 cores		Host power
GPU cluster		200TB	Host power
GPU cluster	4TB	32TB	Host power

Ten TREs reported that Windows (including Windows 10, Windows Server 2012, Windows Server 2019) was the standard build operating system. Four of the participants responded that they could provide both Windows and Linux based on the researcher's request. In one TRE, Ubuntu was the only standard build. From the participants' responses, it was evident that there was great variety in the different specifications available, some having multiple orders of magnitude more capacity than others.

Limitations: While most current researcher needs are met by existing TREs, it is clear from Table 2 that some TREs could find it challenging to support processor intensive projects. Further, most but not all of the TREs gave each project its own isolated virtual machine, which could have implications for isolation and pose an increased security risk if malicious code was able to run and potentially access other projects and their data within a shared system as opposed to a project-specific virtual machine.

Recommendations: TREs should consider the scalability of their infrastructure to support resource intensive projects in future. Use of public cloud infrastructure enables much greater flexibility, for a price, and incorporates robust isolation between virtual machines as standard.

4.2 Data Security Measures Employed in TREs to Mitigate the Risk of Unauthorised Access, Data Loss, and Misuse by Researcher

Unauthorised Access: The participants discussed different measures that were implemented to help prevent unauthorised access. Note, different controls were implemented across the participating TREs, depending on the underlying infrastructure. We present here a full list of the controls that were discussed during the interview although not all the TREs implemented the full list of controls described here.

- Best practice password policy (which would include lock-out after 2 or 3 incorrect attempts)
- Access controls
- Access to TRE only permitted via whitelisted IP addresses
- Fully automated account management
- Sensitive projects may have restrictions on location of researcher (in its strictest form, this could include only permitting access from a specific room (on campus) and via managed devices (restricted machines), or more generally only permitting IP addresses from particular countries)
- Managed file access
- Active Directory hierarchical privileges
- Session recording in place
- Monitoring/audit system such as IBM Guardium, SIEM, Splunk
- Multi-factor authentication
- Network segmentation
- Compartmentalisation to limit access to information to entities on a need-to-know basis to perform certain tasks to protect information
- Multi-vendor firewalls (3 different vendors)
- Patch management
- Bi-annual pen testing

Data Loss: In the TREs, internet access is blocked, and users have limited access rights. The remote access is designed to prevent moving data in and out of the environment, except via the official channels, with appropriate controls in place: virtual hardware ports and copying data out are disabled; some TREs also take steps to impede pasting, though this is not reliably achievable. (The direct paste shortcuts can be disabled, but it is trivial to bypass this with a single command on client systems.) Measures are also in place to detect attempts to export data by other routes. Anti-Malware / Anti-Ransomware software and Data Loss Prevention software (DLP) are used.

Misuse by Researcher: The main countermeasure to misuse by researchers is training. Generally, this reinforces key principles to ensure that researchers understood their responsibilities, and what activities are permitted and not permitted within the environment. Other significant mitigation strategies are checking outputs and reviewing the project scope. In two TREs, researchers must be accredited by a particular organisation before they are granted access to the environment (this accreditation requires the researchers to evidence that they have appropriate qualifications and experience). Furthermore, researchers must sign an investigator's declaration stating that they will not misuse the environment, and line-managers and organisations will be held accountable if a user attempts to do anything malicious.

One TRE uses session recording to help detect misuse. In this TRE, researchers' behaviour such as keystrokes can be monitored. Another TRE uses a monitoring program from Darktrace to detect a user running a tool running on their laptop to take screenshots [37]. Another two TREs

have a full audit log from logon to logoff, and one TRE plans to have logging of activity to enable reconstruction in the event of a breach.

There are many other controls discussed including:

- Researchers are not granted admin access in the TRE
- Researchers only have access to their own project's TRE storage
- Printing, mapping drives, and accessing external drives are not allowed
- Command prompt access disabled
- ISO27001 policy rules via a cloud security posture management system

Recommendations: The above examples of current practices to detect and prevent instances of unauthorised access, data loss, and researcher misuse should be considered by all TREs to further improve security, where appropriate for the specific TRE infrastructure. Furthermore, TREs must have a legal agreement constraining access and usage as their data security measures to mitigate the risk of misuse by the researcher. Monitoring programs to monitor and record researchers' behaviour are also useful to reduce misuse.

4.3 Importing of Data or Code

Participants were asked if they allow researchers to import data or code (including libraries) into the environment, and if so what security measures (e.g., software) are employed to support this process.

- Twelve TREs allow the import of both code and data
- Three allow code (with some restrictions) but not external datasets
- Two allow data but not code
- One allows neither

The import of data or code is subject to gatekeeper approval with a check the import does not contain hidden data, and that the code does not pose a threat to the security of the TRE. This gatekeeper approval process varies between TREs, but typically involves manual checks. In addition to scrutiny of the security risk posed by the data/code, this process could also involve checking file size, file type, magic numbers, and known suffixes. Generally, this process would be supported by virus scans, static code analysis tools, and sample code execution in a sandboxed environment.

Some participants discussed the important role of 'trust', and how training the researchers and trusting that they have no malicious intent is sufficient, based on the low risk of potential damage from malicious code and subject to low sensitivity of the data (see Safe People Section 6 for more detail). Finally, one participant mentioned the role of monitoring to detect any malicious behaviour, so that inappropriate or malfunctioning software would be identified.

Limitations: There was substantial reliance on manual checks to support this process. Further, participants have clear concerns about the security implications of importing malicious code. The main concerns with the process of supporting code or data egress were highlighted as:

1. Ensuring that the AI algorithms or software imported into the environment do not include sensitive data.
2. It can be extremely time consuming for the TRE staff or researchers to manually import code after each small change.

Recommendations: Some of these security concerns could be mitigated by isolating each project within the TRE to minimise potential damage and limiting the privileges of the researchers in the environment, using virtualisation and/or containerisation techniques. Like the recommendations for data checks, there is a clear need for tools that can support the TRE team checking the data and code that researchers wish to bring into the environment. While there

are clear concerns with fully automating this process, developing tools to support these checks could significantly speed up the process and assist with the detection of malicious code.

4.4 Support for Federated Queries of Data from External Sources

Fifteen participants responded that their TRE does not currently support federated queries from external sources, whilst the remaining participants confirmed that their TRE does support this. One of these participants described how their TRE could support federated queries via an integration tool on the Health and Social Care Network (HSCN) facing cloud, using Application Programming Interface (APIs), and Cross-enterprise Document Sharing (XDS) and Image Exchange Portal (IEP) for imaging.

Limitations: Federated queries are difficult to support while maintaining effective privacy and security controls, and not currently available in most TREs at all.

Recommendations: Federated queries enable federated learning that can train ML algorithms from diverse datasets without exchanging data. Federated learning can be effective in diagnosing uncommon diseases, and it can also reinforce data privacy and security if the process of data being stored and processed is supported by privacy-preserving and cryptographic techniques [38, 39, 40]. Further, federated learning complies with data protection regulations including GDPR. However, federated learning is vulnerable to different attacks such as inference attacks (e.g., membership and reconstruction attacks) [41] and poisoning attacks [42, 43] which can violate GDPR. The possibility of these attacks can be mitigated by the application of privacy-preserving mechanisms including Secure Multiparty Computation (SMC), differential privacy, and encrypted transfer learning methods [44]. Supporting federated queries of data from external sources is one feature that is of interest to next generation TREs.

4.5 Audit and Workflow Management

Audit and monitoring are key aspects of a TRE. Many reported that they use project management tools to automate functions such as JIRA [45] which can be customised to record transitions, such as a request being made by a user, review of the data to be exported and subsequent acceptance or rejection of the export. Most of the TREs reported that they keep a copy of exported data.

The level of automation and functionality of auditing differs between TREs. The state of the art includes:

- Real-time alerting on the digital airlock, giving a verbose description of user activity. The reports and alerts generated from this provided the Internet Protocol (IP) address of the user, their username, along with the time, date, file name, file size and some other supplementary fields.
- All activity in the TRE was logged, dashboards were used to support the monitoring of the activity, and reports were automatically generated.
- If a user attempted to take data out that was not permitted, this would be logged. If abnormal patterns were observed then anti-malware (for example, Sophos plus quest tools [46]) would trigger alerts and log tickets on the system. The technical team and data owner would receive an email alert advising them that abnormal patterns had been detected.

Limitations: Many TREs have little or no automation and automated auditing in place, limiting the available reporting and operational insights.

Recommendations: Incorporating a logging and monitoring system into TRE is important. This

system could include login attempts, including username, time and date of access, IP address, the type of activity conducted during the TREs (for example which tools were used, for how long, and any processes running), and details of any imports and exports (including file name, file size, etc.), and the access type (successful or denied). Furthermore, having a real-time alert system can warn the TRE team promptly in case of any malicious attempts and assist in preventing unwanted disclosure and blocking access.

5 Safe Computing

Participants were asked whether their TRE utilises private (on-premise) or public cloud infrastructure. Fourteen of the participants reported that their TREs use a private cloud. There were some concerns from these participants that Data Governance restrictions might make switching to the public cloud difficult. Four of the TREs were already hosted in public clouds and two of the participants reported that they aim to switch to a public cloud in the future. Though costs are generally higher in public clouds, the extra functionality and flexibility make this an attractive option when possible.

6 Safe People

The Safe People controls are measures and policies to ensure that trusted researchers will use the platform in an appropriate manner.

6.1 Controls on the People Who Use the TRE

Best practices for ensuring that the researchers accessing the environment are trustworthy and understand the importance of correct usage of the TRE include: signing legal documents to agree that a researcher would avoid attempting to re-identify any individual; rapid disclosure of any vulnerabilities detected by a researcher; login credentials would be kept private; and notification to the TRE if a researcher was leaving their institution. One participant reported that financial penalties could be a useful deterrent to misuse.

Seventeen of the participants responded that researchers using their TREs are required to complete training. This training typically consists of information governance, GDPR and awareness of issues relating to privacy, ethics, security, information security, Medical Research Council training and SDC. Researchers are typically required to complete the training annually, or prior to the start of each project. The nature of this training and subsequent contract or terms-of-use are typically determined by the data owner. For example, government security clearance is requested by Defence Science and Technology Laboratory (DSTL) for access to their data.

In fifteen TREs, researchers sign an agreement not to misuse the environment or the data. This agreement is also signed by a senior member within each organisation. One participant stated that if a researcher is a student, a supervisor needs to sign the agreement too. There were a range of penalties applied across the TREs for violating the user-agreement, which in the most extreme form could result in job loss, and disciplinary measures, or in some cases compulsory retraining. Project approval is also required by the relevant data controller, and in some cases the project also has to be signed off by an ethics committee. According to one TRE, conditions specified in a Non-Disclosure Agreement or Access Request Form would impose constraints regarding appropriate use of the data and could pass all responsibilities for ensuring the data was being used correctly on to the sponsoring organisation.

Recommendations: Training such as Information Governance training is vital to ensure researchers understand their responsibilities and should be considered by all TREs. Via training, researchers will clearly understand what they are allowed to do with the data. TREs must implement suitable review and management processes to further ensure that researchers are using the TREs appropriately.

6.2 Controlling Access to the Environment for Trusted Users Only

Fourteen participants stated that access to their TREs is limited to those researchers associated with an approved (trusted) organisation. Further, two of these participants stated that access was limited to organisations in the same country as the TRE, as specified by the data custodian. For one of them, commercial organisations were not allowed to access the TRE under any circumstances. For other TREs who did permit commercial organisations to access the environment, the criteria for approving these organisations were generally set higher than other organisations (e.g., universities). In one TRE, although requests from commercial organisations were considered, they needed a university sponsor or health sponsor to be approved. Another participant responded that commercial customers did not need to be associated with an academic institution. In this case, a review committee considered which projects would be approved for commercial customers.

In one TRE, access is granted only to their own university's users. In this TRE, an external visitor account would only be granted access if the visitor was sponsored by university staff. In another TRE, researchers could access the environment from a university, or NHS based organisation (i.e., using whitelist IP addresses). One TRE adopted additional restrictions for the researchers, for example ensuring that access was only permitted from a safe room, or that the device used to access the TRE was a managed device and not a personal device. In this instance, these restrictions were set by the data controller.

7 Participant Recommendations for TRE Enhancements

Six participants indicated that they would like to improve support for programming capabilities in the environment, for example Python and R, to advance the analytical capabilities of their TRE and subsequently support mass-scale studies. Support for bringing data, algorithms, and code into the environment was frequently described as another high priority feature. However, the licensing of proprietary software tools presents a further limitation with regards to bringing software into the environment, as not all licences cover usage within a TRE.

Two participants confirmed that they would like to support federated learning to advance data movement among TREs, where data sets need to be shared and accessible. Support for additional data modalities, such as imaging and genomic data, needs to follow a proper risk assessment and TREs would have to ensure that they liaise with data custodians regarding the specific risks. It was widely acknowledged by the participants that there were many security challenges around allowing researchers to bring their own data and code into the environment, and until solutions to these challenges have been developed, many TREs will be reluctant to support this.

Five participants indicated that they would like to simplify the process for researchers to access data within their TRE in the future. The process for this, and the checks required before researchers are granted access to the data were perceived as cumbersome and slow. Sometimes, this administrative process is further delayed due to backlog of project review requests, committees being slow to make decisions, ethics board approval, getting researchers

onto relevant training courses and privacy training. Researchers are eager to have access to the TRE and its data promptly, hence there was a desire from TREs to simplify this process.

Eight participants discussed how they would like to improve the governance processes. One of the participants stated that all datasets in their TRE were treated as high risk, and had to go through the same governance process, although some data sets were actually low risk. One participant suggested that it would be useful to conduct a national risk-benefit analysis of sharing standardised data sets for research. The participant acknowledged that was no systematic approach to review data sets to determine if there were certain conditions under which these could be used by researchers without the full governance checks.

Some TREs are considering migrating to the public cloud for improved scalability and flexibility, including GPU access and greater on-demand computing power and reduced management overheads, while several have already made this transition.

One TRE is looking to enhance the security of their TRE through improved logging of activities such as data copying between machines, and better behaviour tracking.

Finally, one of the participants discussed concerns with Intellectual Property (IP) when code was developed within the TRE. The participant acknowledged that researchers may have concerns regarding how the code that they develop or test in a TRE could be accessed by the TRE operators. Policies and practices to governing this should be in place to protect both parties. Technical solutions to this such as Trusted Computing and enclave approaches could also be explored.

8 Conclusion and Future Work

This research reviewed the existing controls employed by UK and international TREs who participated in our structured interviews. These controls cover a subset of the 7 “Safes”, comprising Safe people, Safe setting, Safe computing (an extension of Safe setting), Safe data and Safe outputs. The features that are significant to improve for next generation TREs are:

- Advancing analytical power (High Performance Computing Clusters and GPUs) available within the environment to support large-scale studies
- Bringing data/algorithms/code into the environment, and addressing the security challenges arising from this
- Being able to develop ML and AI algorithms within the TRE and export these
- Supporting federated queries of data from external sources
- Support for additional data modalities such as imaging and genomic data
- Simplifying the process of accessing data for researchers
- Scalability

The paper analysed the extent to which TREs are able to support the import and export of different data types. The process used is largely manual, with some TREs making use of software to support this process. Finding suitable software to support automation of the DC process was identified as a key priority for most of the TREs. Further, the application of ML techniques in TREs could be useful in predicting the malicious use of accessed data by researchers. It was evident that in most TREs, there are no specific tools to manage and mitigate the potential risk of re-identification, and it relies on analysts’ knowledge, judgement, and communication with the data controller.

There is a lack of support for AI and ML development in TREs and concern that researchers could perform malicious activities due to the AI and ML structure. For instance, exporting sensitive data that could be vulnerable to exposure following attacks against the AI model, or over-training the AI algorithm. The difficulties in detecting these exports acknowledged as a

significant challenge by the participants.

The computing power available to researchers is generally adequate for current needs (observational studies using statistical analysis tools), although there was a clear desire to ensure that this was scalable to meet researcher requirements for analysing big data and for AI development. Some TREs already appear significantly constrained. There is significant variety in the extent of the security measures employed to mitigate the risk of unauthorised access, data loss, and misuse by the researcher, and some concerns regarding the implications of next generation capabilities on the security of the TRE and protecting the data. Furthermore, there is the need for advanced Information Governance for TREs encompassing incoming and outgoing automated data feeds, ad hoc incoming data and algorithms, and ad hoc outgoing data and algorithms. Finding appropriate solutions to address these improvements should be explored in future work.

Funding

This project was supported by MRC and EPSRC [grant number MR/S010351/1] programme grant: Interdisciplinary Collaboration for efficient and effective Use of clinical images in big data health care REsearch: PICTURES [grant number MR/S010351/1].

This work was also supported by Health Data Research UK (HDR UK: 636000/ RA4624) which receives its funding from HDR UK Ltd (HDR-5012) funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation (BHF) and the Wellcome Trust.

Abbreviations

References

1. The Scottish Government. A Charter for Safe Havens in Scotland. 2015. URL: <https://www.gov.scot/publications/charter-safe-havens-scotland-handling-unconsented-data-national-health-service-patient-records-support-research-statistics/> [accessed Sep 15, 2021]
2. Hubbard T, Reilly G, Varma S, & Seymour D. (2020). Trusted Research Environments (TRE) Green Paper (2.0.0). Zenodo. [doi: 10.5281/zenodo.4594704]
3. Institute of Medicine; Board on Health Sciences Policy; Committee on Strategies for Responsible Sharing of Clinical Trial Data. Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk. Washington (DC): The National Academies Press (US). April 20, 2015.
4. Lea N, Nicholls J, Dobbs Ch, Sethi N, Cunningham J, Ainsworth J, Heaven M, Peacock T, Peacock A, Jones K, Laurie G, Kalra D. Data Safe Havens and Trust: Toward a Common Understanding of Trusted Research Platforms for Governing Secure and Ethical Health Research. JMIR Med Inform. 2016 Jun 21;4(2):e22. PMID: 27329087. doi:10.2196/medinform.5571
5. University of Dundee. HIC Safe Haven. Available from: <https://www.dundee.ac.uk/hic/hicsafehaven/> [accessed Sep 16, 2021].
6. The University of Edinburgh. eDRIS: enabling research access to Scottish health datasets. 2020. URL: <https://www.ed.ac.uk/edinburgh-international-data-facility/updates-events/electronic-data-research-and-innovation-service> [accessed Sep 16, 2021].
7. Jones KH, Ford DV, Thompson S, Lyons RA. A Profile of the SAIL Databank on the UK Secure

- Research Platform. *International Journal of Population Data Science*. 2019; 4(2). doi: <https://doi.org/10.23889/ijpds.v4i2.1134>
8. Jones KH, Ford DV, Ellwood- Thompson S, Anthony Lyons R. The UK Secure eResearch Platform for public health research: a case study. *The Lancet*. 2016 Nov; 388(S62). doi: [https://doi.org/10.1016/S0140-6736\(16\)32298-X](https://doi.org/10.1016/S0140-6736(16)32298-X)
9. ADR UK. Office for national statistics. <https://www.adruk.org/about-us/our-partnership/office-for-national-statistics/> [accessed Sep 16, 2021].
10. ADR UK. Administrative data is invaluable resource for public good let's use it. Available from: <https://www.adruk.org/> [accessed Sep 16, 2021].
11. UK Data Service. Access levels and conditions. Available from: <https://www.ukdataservice.ac.uk/use-data/secure-lab.aspx> [accessed Sep 16, 2021].
12. Desai T, Ritchie F, Welpton R. Five Safes: designing data access for research. *Economics Working Paper Series 1601*. 2016. <https://www2.uwe.ac.uk/faculties/bbs/Documents/1601.pdf> [accessed Apr 6, 2020].
13. El Emam Kh, Dankar F K, Protecting Privacy Using k-Anonymity. *Journal of the American Medical Informatics Association*. 2008; 15(5): 627–637. doi:10.1197/jamia.M2716
14. Templ M, Kowarik A, Meindl B. Statistical Disclosure Control for Micro-Data Using the R Package sdcMicro. *Journal of Statistical Software*. 2015; 67(4). doi: 10.18637/jss.v067.i04.
15. PRIVITAR. Available from: <https://www.privitar.com/> [accessed June 10, 2021]
16. TriNetX. TriNetX, InSite Unite to Establish World's Largest Clinical Research Network. Apr 2, 2019. <https://trinetx.com/inSITE/> [accessed Sep 16, 2021]
17. Rushton S. Privacy Analytics: It's nothing personal. UOttawa. Available from: <https://research.uottawa.ca/perspectives/privacy-analytics-its-nothing-personal> [accessed Sep 16, 2021].
18. FUTURE OF PRIVACY FORUM. De-identification. <https://fpf.org/issue/deid/>. [accessed Sep 16, 2021].
19. Hogue P. The Risk of Triangulation: You May Just be a Piece of the Puzzle. SECURITYWEEK. September 11, 2018. <https://www.securityweek.com/risk-triangulation-you-may-just-be-piece-puzzle> [accessed Sep 16, 2021].
20. Guru J, Damecha H. Digital watermarking classification: a survey. *International Journal of Computer Science Trends and Technology (IJCSST)*. 2014 Sep-Oct ;2(5): 8-13.
21. European Medicines Agency. Data anonymisation - a key enabler for clinical data sharing. December 4, 2018 . https://www.ema.europa.eu/en/documents/report/report-data-anonymisation-key-enabler-clinical-data-sharing_en.pdf [accessed Sep 16, 2021].
22. Rigaki M, Garcia S. (2020). A survey of privacy attacks in machine learning. ArXiv. Preprint posted online on July 15, 2020. <https://arxiv.org/pdf/2007.07646.pdf>
23. Fredrikson M, Jha S, Ristenpart Th. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*; 2015 Oct: 1322–1333. doi: <https://doi.org/10.1145/2810103.2813677>
24. Nind T, Sutherland J, McAllister G, Hardy D, Hume A, MacLeod R, Caldwell J, Krueger S, Tramma L, Teviotdale R, Abdelatif M, Gillen K, Ward J, Scobbie D, Baillie I, Brooks A, Prodan B, Kerr W, Sloan-Murphy D, Herrera J, McManus D, Morris C, Sinclair C, Baxter R, Parsons M, Morris A, Jefferson E. An extensible big data software architecture managing a research resource of real-world clinical radiology data linked to other health data from the whole Scottish population. *Giga Science*. 2020 Oct; 9(10) .doi: 10.1093/gigascience/giaa095

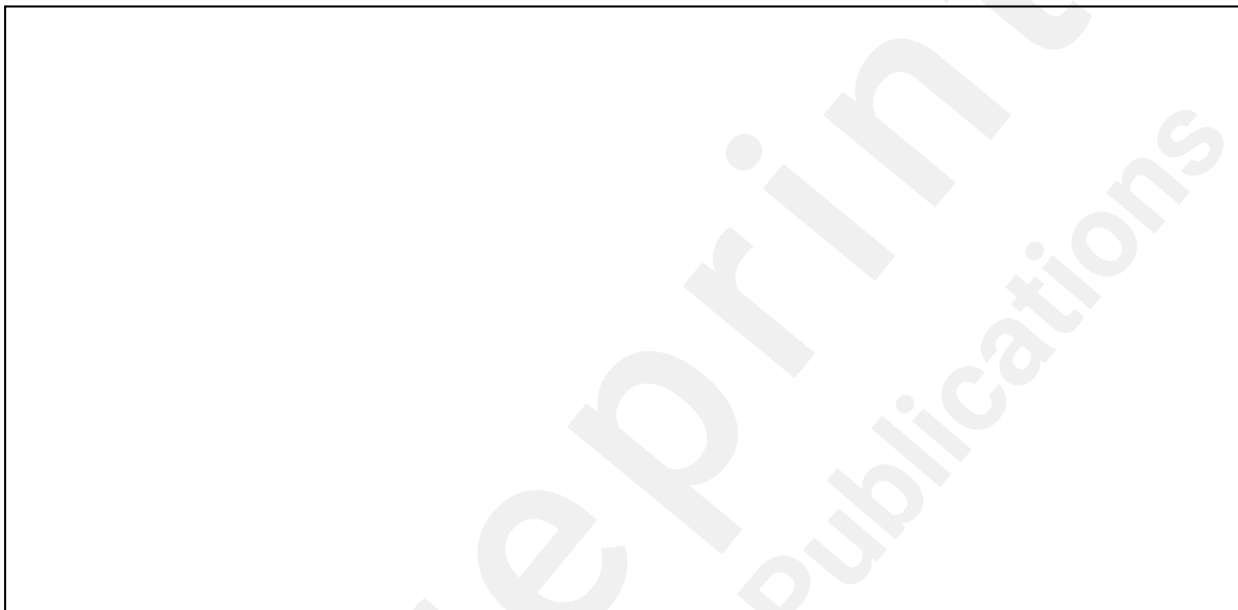
25. Nicolae MI, Sinn M, Tran MN, Buesser B, Rawat A, Wistuba M, Zantedeschi V, Baracaldo N, Chen B, Ludwig H, Molloy IM, Edwards B. Adversarial Robustness Toolbox v1.0.0. ArXiv. November 15, 2019. <https://arxiv.org/abs/1807.01069>
26. Liu Y, Wen R, HeX, Salem A, Zhang Z, Backes M, De Cristofaro E, Fritz M, Zhang Y. ML-Doctor: Holistic Risk Assessment of Inference Attacks Against Machine Learning Models. February 2021. https://www.researchgate.net/publication/349045039_ML-Doctor_Holistic_Risk_Assessment_of_Inference_Attacks_Against_Machine_Learning_Models
27. Shokri R, Stronati M, Song C, Shmatikov V. Membership inference attacks against machine learning models. In 2017 IEEE Symposium on Security and Privacy; 2017 May 22-26; San Jose, CA, USA:. doi: [10.1109/SP.2017.41](https://doi.org/10.1109/SP.2017.41)
28. Salem A, Zhang Y, Humbert M, Berrang P, Fritz M, Backes M. ML-leaks: model and data independent membership inference attacks and defenses on machine learning models. Proceedings of the 26th Annual Network and Distributed System Security Symposium (NDSS); 2019; San Diego, California, USA.
29. Qian J, Li XY, Zhang CH, Chen L. De-anonymizing social networks and inferring private attributes using knowledge graphs. In 35th Annual IEEE International Conference on Computer Communications; 2016 Apr 10-14; San Francisco, CA, USA.
30. Al-Rubaie M, Chang JM. Reconstruction attacks against mobile based continuous authentication systems in the cloud. IEEE Transaction on Information Forensics and Security. 2016; 11(12).
31. Takemura T, Yanai N, Fujiwara T. Model extraction attacks against recurrent neural networks. Journal of Information Processing. 2020; 28: 1010-1024. doi: <https://doi.org/10.2197/ipsjip.28.1010>
32. Reith RN, Schneider T, Tkachenko O. Efficiently stealing your machine learning models. In Proceedings of the 18th ACM Workshop on Privacy in the Electronic Society; 2019; London, UK.
33. Zhang Y, Jia R, Pei H, Wang W, Li B, Song D. (2020). The secret revealer: Generative model-inversion attacks against deep neural networks. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR) 2020; Seattle, WA, USA.
34. Aslanyan Z, Vasilikos P. Privacy-preserving Machine Learning. Whitepaper: A Practical Guide. 2020. Available from: <https://www.alexandra.dk/wp-content/uploads/2020/10/Alexandra-Instituttet-whitepaper-Privacy-Preserving-Machine-Learning-A-Practical-Guide.pdf>
35. Macrae C. (2019). Governing the safety of artificial intelligence in healthcare. BMJ quality & safety. 2019; 28(6). doi: <http://dx.doi.org/10.1136/bmjqs-2019-009484>
36. Paprica PA, Sutherland E, Smith A, Brudno M, Cartagena RG, Crichlow M, Courtney B, Loken CH, McGrail KM, Ryan A, Schull MJ, Thorogood A, Virtanen C, Yang K. Essential requirements for establishing and operating data trusts: practical guidance co-developed by representatives from fifteen Canadian organizations and initiatives. International Journal of Population Data Science. 2020; 5(1). doi: <https://doi.org/10.23889/ijpds.v5i1.1353>.
37. DARKTRACE. Featured Resources. <https://www.darktrace.com/en/resources/> [accessed Sep 16, 2021].
38. Phong LT, Aono Y, Hayashi T, Wang L, Moriai SH. Privacy-Preserving Deep Learning via Additively Homomorphic Encryption. IEEE Transactions on Information Forensics and Security. 2018; 13 (5). doi: [10.1109/TIFS.2017.2787987](https://doi.org/10.1109/TIFS.2017.2787987)
39. Wei K, Li J, Ding M, Ma C, Yang H, Farokhi F, Jin S, Quek T, Poor V. Federated Learning with Differential Privacy: Algorithms and Performance Analysis. IEEE Transactions on Information Forensics and Security. 2020 Apr; 15: 3454-3469. doi: [10.1109/TIFS.2020.2988575](https://doi.org/10.1109/TIFS.2020.2988575)
40. Truong N, Sun K, Wang S, Guitton F, Guo Y. Privacy preservation in federated learning: An insightful survey from the GDPR perspective. COMPUTERS & SECURITY. 2021; 110.

41. Dwork C, Smith A, Steinke T, Ullman J. Exposed! A Survey of Attacks on Private Data. *Annu Rev Stat Appl.* 2017; 4(1): 61-84.
42. Bhagoji . AN, Chakraborty S, Mittal P, Calo S. Analysing Federated Learning through an Adversarial Lens. *International Conference on Machine Learning.* 2019; 634-643.
43. Bagdasaryan E, Veit A, Hua Y, Estrin D, Shmatikov V. How to backdoor federated learning. *International Conference on Artificial Intelligence and Statistics.* 2020: 2938-2948.
44. Salem M, Taheri S, Yuan JS. Utilizing Transfer Learning and Homomorphic Encryption in a Privacy Preserving and Secure Biometric Recognition System. *Computers.* 2019; 8(1). doi: <https://doi.org/10.3390/computers8010003>
45. ATASSIAN. Jira Software. <https://www.atlassian.com/software/jira> [accessed Sep 16, 2021].
46. SOPOS HOME. Security and Privacy for the entire family. <https://home.sophos.com/en-us.aspx> [accessed Sep 16, 2021].

Appendix A – Interview Questions

1) Background Information:

- a. Name of the organisation that you work for:
- b. Name of the Data Safe Haven (DSH) that you work for:
- c. Job title:
- d. Brief explanation of your role in the DSH:
- e. Can you describe what works well in your DSH?
- f. Can you describe what features you would like to see in place to improve the DSH?



Safe Outputs

2) Select what types of data a researcher can export from the DSH:

- Aggregate level graphs and tables
- Individual level, anonymised data
- Weights of an AI algorithm
- Weights and code for an AI algorithm
- Software source developed within the DSH
- Software executable developed within the DSH
- Other types of data – please list



3)

- a) Do you receive requests to export other data types? If yes please describe.
- b) Do you have any plans to enable export of any additional types of data in the future? If yes please describe.

- 4)** Describe the process you use for checking disclosure control on data to be exported. Please cover
- a) what software you use or consider for this purpose
 - b) are there any issues (eg cost, rate of false positives) that have prevented suitable software being adopted
 - c) when the data can be exported (e.g. once, daily)
 - d) how that data can be exported
 - e) what restrictions do you impose on this process?

- 5)** If software is used for disclosure control (as described in question 4), what checks does this software perform?

6) What manual checks do you use for disclosure control?



- 7)
- a) Have any known breaches or near-miss incidents occurred? (For example, was a spreadsheet file cleared for export without realising there was additional data stored within the Undo history?)
 - b) Are you aware of any potential gaps in your disclosure control checks?(For example, could a researcher change the colour of text to the same as the background colour of a file to be exported and this be missed by the manual checks?)



8) What is the minimum number of individual data points allowed within a cell to be exported? For example, a cell count within a table has to be >5 individuals.



9 Safe Data

- 9) As part of strengthening safeguards, which tools and techniques are used in your DSH to manage and reduce the potential risk of re-identification?

- 10) Digital watermarking is a technology in which identification information is embedded into the data carrier in ways that cannot be easily noticed, and in which the data usage will not be affected. This technology often protects the copyright of multimedia data and protects databases and text files. Digital watermarking can be effectively used to trace disclosure of health data. Do you think watermarking can be used to preserve health data privacy when data is disclosed to researchers via DSH?

- Yes
- No

If yes, have you used this approach on your developed DSH? Describe your approach.

10 Safe Setting

11)

- a) What computing power do you offer a “standard” DSH user (CPUs, GPUs, memory and storage)?
- b) What OS is used on your “standard” build (Windows/Linux/other)?
- c) What environment do you offer if a VM is not used? (For example, Amazon SageMaker for building machine learning pipelines in a web interface)

- 12)** What is the maximum computing power you offer DSH “power” users (CPUs, GPUs, memory and storage)?

- 13)** What are the data security measures that are employed in your DSH to mitigate the risk of the following?
- a) Unauthorized access
 - b) Data loss
 - c) Misuse by researcher

- 14)** Do you have a standard/base build for the VMs?
- Yes
 - No

If yes, what OS and tools are installed on the VM?

15)

a) Do you implement internal isolation between projects/users/VMs? If yes, how is this managed?

b) What measures do you employ to ensure that researchers cannot execute malicious code in the environment?



16) What measures do you employ to control any external access to the VM (e.g. USB, connecting external drives, connecting to the internet)?



17)

a) What security checks do you employ on the VM?

b) Do you have internal red team/testers that check the security?



18) Do you allow the researchers to have a custom-built environment?

If yes, what security measures do you employ to check the custom VM?

19) Do you allow researchers to modify the environment at a later stage?

If yes, what security measures do you employ to check the modified environment?

20) Are researchers allowed to import data or code (including libraries) into the environment?

If yes, what security measures do you employ at this stage? For examples how do you scan imported software/tools to ensure that they will not compromise the security and integrity of the DSH?

21) Does your DSH support federated queries of data from external sources? If so, please give details.

1

11

12 Safe Computing (an extension of Safe setting)

22) Is your DSH a private or public cloud? Please provide details.



13 Safe People

23) What controls do you put on the people who use the DSH? For example, data governance training, signing a legal document with terms and conditions of use.



24) Does access to the environment have to be via a recognised “trusted” organisation i.e. from a university network within the UK? Please provide details.



25) Can the environment be accessed from anywhere in the world? If not, please provide specifics.

Preprint
JMIR Publications

Evaluating the functionality of the DSH

Based on your experience with DSH, please indicate the extent to which you agree (or disagree) with the following statements:

Statement	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
DSH is a dynamic instrument that can contribute to future developments in the science, technology, and practices of genomic and health-related data sharing.					
DSH can serve as a tool for the evaluation of responsible research by research ethics committees and data access committees.					
Information security and privacy controls that are applied to DSH can be applied effectively to all operations, services and systems that process sensitive data.					
Anonymization of Personally Identifiable Information (PII) or Protected Health Information (PHI) on DSH can preserve data privacy and mitigate the possibility of re-identification and information leakage.					
Application of machine learning techniques on DSH can be useful in predicting the malicious use of accessed data by researchers.					
Identifiers, recognisable identifiers, and sensitive attributes must be removed completely.					

DSH owner can fully ensure that the researcher only uses the data appropriately.					
There is no concern that accessing data in the DSH may be damaging to the patient.					
Statistical results must not be disclosed in the public domain.					

14 Additional Questions If Time Allows

- 26) Which of the following are important for you to consider within your DSH?
- Data privacy
 - Fine-grained access control
 - Efficiency
 - Scalability
 - Ability to export data
 - Ability to export algorithms
 - Ability to import existing algorithms from external storage
- 27) Does your organisation have any policies or principles governing the safe use of data in your DSH?
- Yes
 - No
- If yes, are these policies or principles influenced by any standards, laws of government polices?
- Yes
 - No
- If yes, can you list them?
- 28) Is your DSH approved / validated? If so by which organisation?
- 29) In your opinion, which of the following principles do you think is the most important to ensure the responsible sharing of genomic and health-related data?
- Respect Individuals, Families and Communities
 - Advance Research and Scientific Knowledge
 - Promote Health, Wellbeing and the Fair Distribution of Benefits
 - Foster Trust, Integrity and Reciprocity
- 30) Who is responsible for managing your data security and privacy program?

Preprint
JMIR Publications