# TREEHOOSE: Trusted Research Environment and Enclave for Hosting Open Original Science Exploration

# Final Report

## 1. Background

Trusted research environments (TREs) are necessary for the ethical and secure management of sensitive data. Designing and operating a TRE in the cloud requires considerable custom work, with a challenging learning curve for operations staff. The implementation, maintenance and development of TREs requires unique skillsets for which there is a current shortage. Upskilling of new and existing staff is crucial for the continued growth and development of Data Science in the UK.

The Health Informatics Centre (HIC) have been running a TRE for over a decade and have recently collaborated with AWS to develop a Cloud TRE which is now in production. The TREEHOOSE proposal was to create and share an open-source infrastructure-as-code implementation, with documentation, for both operating and performing research within a modern cloud TRE. The aim of the open source architecture was to

- Reduce the learning curve for TRE operators and users alike.
- Reduce the cost of migration to the cloud for other TREs.
- Aid portability of code between TREs.
- Make federation more straightforward through the use of common cloud deployments.

The TREEHOOSE project also upskilled project team members for the benefit of future research projects.

None of this work is possible nor sensible without engagement with the public and patients as they are the final arbiters of acceptable reuse of their data. TREEHOOSE ran workshops on the trustworthiness of cloud computing with research data and generated reports on our findings.

## 2. Outputs

The project’s three work packages each had deliverables which were met.

### 2.1. Work Package 1 – Open Source TRE

In this WP the main aim was to deliver an open source TRE based on the HIC TRE. In developing the project proposal and in its initial phase we identified an AWS internal project building a TRE. The HIC TRE had the benefit of HIC’s over ten years’ experience of running and maintaining a TRE meaning it fit very well the typical needs of health data researchers and health informaticians. The TRE codebase had diverged quite significantly from the upstream Service Workbench (SWB) upon which it was built. The AWS TRE had the advantage of being developed in conjunction with the Service Workbench infrastructure team thereby ensuring compatibility with the upstream codebase. It, however, lacked several of the features of the HIC TRE and was not adapted to the typical use-cases of clinical researchers.

It was planned to merge the HIC and AWS TRE codebases to produce the Open Source TREEHOOSE codebase. The planning process took two months of carefully working through the two codebases to develop a “backlog” of issues and features which needed to be migrated from one or other into the new TREEHOOSE project.

Once the planning stage was completed, an agile process was implemented with two-weekly sprints and stand-ups twice a week managed via an AWS project manager and involving a team of four AWS developers and two members of the HIC team, one infrastructure and one research software engineer.

WP1 was structured into four sections: Architecture, Security, Deployment and Documentation. Each is described in more detail below.

### 2.1.1. Architecture

The fundamental building blocks of the TREEHOOSE TRE on AWS infrastructure are SWB, the Data Lake and AWS Control Tower. SWB is an existing open-source application for managing project workspaces in the cloud for researchers without them needing to have a deep understanding of how cloud computing infrastructure works, thereby making the experience more streamlined and user friendly. SWB contains many of the features required for running a TRE, but requires careful customisation to have the necessary restrictions required by a TRE. AWS Control Tower helps manage the security of an organisation’s AWS cloud accounts, and is an important component of the required security layers ensuring that the governance requirements for running TRE are met. Finally, data which are used within a TRE must be stored and made available in the cloud which is where the third building block, the Data Lake, comes in. All data managed by the TRE maintainers are stored within the Data Lake and then made available only to users or projects with the appropriate permissions. A Data Lake may be managed by an outside organisation, allowing them to retain full control of their data.

Figure 1 shows a schematic of the TREEHOOSE architecture. In the centre is an example project which is managed by the IT administrators and Data Managers, on the left. The project has access to data from the Data Lake for which it has been granted permission, and is also managed by the Data Managers. On the right the researcher gets secure access to the project, data and computing resources - via SWB - as defined by the IT and Data Managers. It is important to highlight in Figure 1 the interaction between the different services and the user personas which manage the proper running of the TRE and the available functionality within it. The TRE can be optimised to suit differing models as required by local policies or regulations.

TREEHOOSE includes three additional functions which although optional form an important part of managing a TRE and making it useful for research: data egress, workspace backup and budget controls. Although a TRE is secure by definition and needs to comply with the 5 safes, researchers do need to be able save some results of their work outside a TRE in order to publish academic papers, write reports or to allow others to review their work for transparency and validation. Science is dependent on being able to build upon other’s work. The process of saving results outside of TREs is called “data egress”. In doing so a request is made to export one or more files which is manually verified by members of TRE data or governance teams – as required by local policies or regulations – to meet the strict limits of disclosure control. If the files are approved, they are provided to the researcher via a non-TRE mechanism (e.g. email) and they are free to disseminate them further. Egressing data is a common and important task for TREs to support.

The second optional functionality is a backup mechanism to enable the saving of project workspaces and data. The backup mechanism available in TREEHOOSE enables TRE administrators to backup researcher workspaces in case of errors. The backups work for both EC2-based workspaces as well as SageMaker notebooks.

A common concern regarding cloud-based computing workflows is the open-ended nature of the costs. With a traditional local infrastructure computing hardware is purchased up-front as capital cost, whereas with cloud the cost is a consumable cost for as long as the workflow is running. This means the costs in the cloud are less predictable. Thus, a system of controls has been made available as a final, optional component in TREEHOOSE to allow administrators to monitor spending, automatically send reports and create actions to avoid over-spending given a fixed budget.

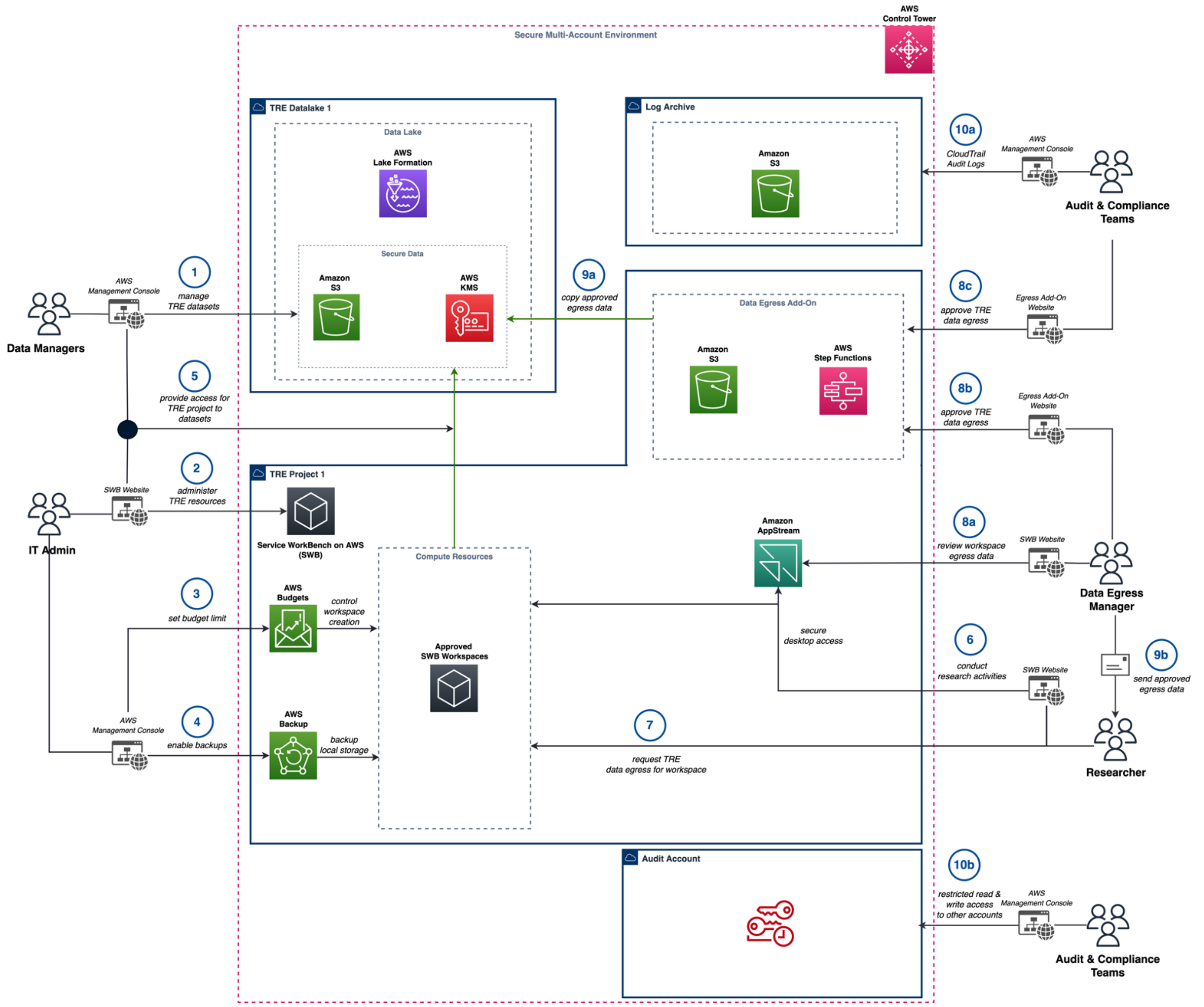

*Figure 1. The TREEHOOSE Architecture. AWS services and their interactions with the different user personas are shown.*

### 2.1.2. Security

Security in the AWS cloud follows a shared responsibility model where AWS, as providers of the cloud services, together with the customer, who runs the TRE, both share responsibility for securing the environment.

The core security is regarding data as they are the single most sensitive entities in a TRE to security concerns. A TRE will be storing and making available to authorised persons data containing sensitive and/or personal information regarding individuals or private organisations. Data within the data lake is encrypted during transfer and at rest ensuring that unauthorised entities cannot see the data.

Only approved users, complying with the "safe people" requirement of a TRE, are allowed access. The TREEHOOSE TRE enables the use of secure user management systems that integrate with many existing solutions, which means it can work with processes that are already in place at host organisations without adding extra complications. User access is always under the control of the TRE manager who can add and remove users as required.

### 2.1.3. Deployment

One of the key benefits of TREEHOOSE has been the ease of deployment. Now, with a bit of pre-existing knowledge, a TRE can be deployed by an organisation on their own in a matter of days. Previously, this required very detailed knowledge of cloud infrastructure as well as the recruitment of specialist services to create a TRE on an almost ad hoc manner. The result of this is that up to now most TREs are quite different to each other and have few shared systems.

The deployment is performed via infrastructure-as-code to automate almost all of the process, with a few manual steps required by the TRE administrators. We have created template code, in CloudFormation, which can be modified to suit particular TRE use cases and which incorporate the mandatory and optional components mentioned in the Architecture section. TREEHOOSE provides a common base and also flexibility to TRE managers to tailor the software to best meet their needs.

### 2.1.4. Documentation

Setting up, running and maintaining a TRE is not a trivial exercise. Experience in being comfortable with the technical and governance requirements which make a TRE work matters substantially. TREEHOOSE encompasses the >10 years' experience from the team at HIC in running a TRE for supporting research on NHS clinical data from two Scottish Health Boards. It is important to stress that TREEHOOSE is a technical solution which makes the other requirements for TREs easier to manage, but it does not replace the need for good governance and administrative processes.

In making the infrastructure-as-code in TREEHOOSE open-source we have made a big step forward in enabling more consistency across TREs and lowered the bar to access for organisations. However, on its own this is not enough which is why we have developed a substantial "Operations Guide" together with the code to support administrators and managers to set up, run and maintain a TREEHOOSE installation.

The documentation covers all aspects of the architecture shown in Figure 1 including the optional components. We have included over 3,000 lines of text and over 75 images to explain all steps and processes.

The codebase was made available on GitHub at https://github.com/HicResearch/TREEHOOSE/ as v1.0.0-beta (10.5281/zenodo.6908253).

The TREEHOOSE team have demonstrated the TRE environment to several groups and organisations who have shown great interest in the architecture for their use and wish to collaborate on developing new features. As far as we know 2-3 groups have downloaded and "spun up" the TREEHOOSE TRE by themselves, which is a significant improvement over the AWS TRE solution that requires a two-week engagement with their consultants. We are now a partner of the Smart Manufacturing Data Hub- a multi-centre consortium to support SMEs to access a secure environment for data analysis and data sharing. TREEHOOSE is the secure environment chosen for this 3-year project.

## 2.2. Work Package 2 – Secure Enclave Proof-of-Concept

The HIC TRE is used to manage safe and secure access to patient health records from NHS Tayside and NHS Fife. There is an opportunity for the same infrastructure to be used to support research disciplines across different sectors in line with DARE UK (Data and Analytics Research Environments UK)'s aims. In many commercial sectors, intellectual property (IP) concerns are paramount in developing novel technology or knowledge for competitive advantage.

A TRE is a safe and secure infrastructure for data owners to be able to provide access to third parties, however, it does not address the concerns of the third parties sharing IP with the TRE. An example is an organisation that has developed a novel machine learning (ML) algorithm on an existing dataset and now wishes to validate the performance on a new dataset which they previously did not have access to. They would like to use a TRE, but also need to protect their IP. A common solution is to provide an anonymised and minimal version of the data to the third party via a data sharing agreement (DSA) where how and for how long the data are used is stipulated. This is potentially unsatisfactory for both parties as the data owner loses control of their data and the algorithm developers may compromise their needs in order to comply with the DSA. There has been public criticism of this type of DSA in particular with large multinational corporations.

There is another option which enhances the security of sensitive data and protects the IP requirements of third parties. A secure enclave is a cryptographically secure environment which runs a single process and access is controlled by whoever has ownership of the encryption keys.

The aim of WP2 in TREEHOOSE was to develop a secure enclave capability within the open-source TRE to enable the running of an example algorithm, in a cryptographically secure environment inside the TRE, whilst securing the IP rights of the algorithm developer. From an existing collaboration with the PICTURES project we obtained a ML algorithm (available at https://github.com/HicResearch/TREEHOOSE-Enclave-PoC) which predicts whether brain MRI scans were performed as T1 or T2. Data from the publicly-available IXI dataset – provided by Imperial College London – was used as a Proof-of-Concept (PoC) for incorporation into the secure enclave using the AWS Nitro Enclavehttps://aws.amazon.com/ec2/nitro/ service. The Nitro Enclave service was adapted in two ways for use within a TRE; 1) the container image was changed to be able to incorporate ML code within the enclave, and 2) be able to communicate to the AWS Key Management Service (KMS) securely from within the restricted networking environment of a TRE. Both the features were novel and we believe are unique. Figure 2 shows a schematic of how the secure enclave has been incorporated into the TREEHOOSE architecture. The upper section is under the control of the external user where they have their own AWS account which is used to add their algorithm into the secure enclave. The middle and lower sections are part of the TRE. Only three arrows cross the TRE boundary: two are the enclave and encrypted code which are ingressed manually by TRE managers, and one is a message to KMS to allow decryption of the enclave during run time. The message can only access a known endpoint and contains no data.

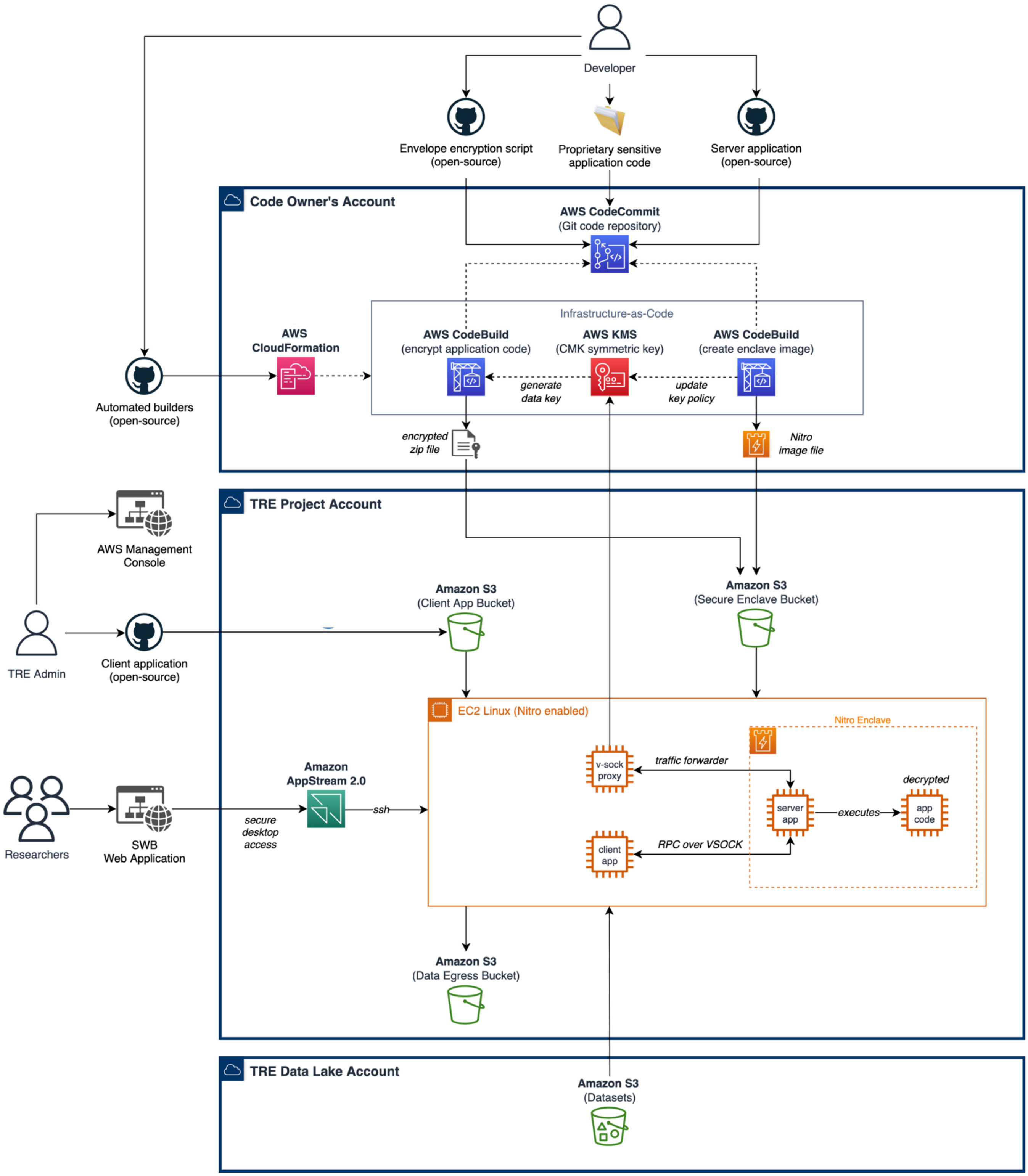
Developer
Envelope encryption script (open-source)
Proprietary sensitive application code
Server application (open-source)
Code Owner's Account
AWS CodeCommit (Git code repository)
Infrastructure-as-Code
AWS CloudFormation
AWS CodeBuild (encrypt application code)
AWS KMS (CMK symmetric key)
AWS CodeBuild (create enclave image)
generate data key
update key policy
encrypted zip file
Nitro image file
Automated builders (open-source)
TRE Project Account
AWS Management Console
TRE Admin
Client application (open-source)
Amazon S3 (Client App Bucket)
Amazon S3 (Secure Enclave Bucket)
EC2 Linux (Nitro enabled)
Nitro Enclave
v-sock proxy
traffic forwarder
server app
executes
decrypted app code
client app
RPC over VSOCK
Researchers
SWB Web Application
secure desktop access
Amazon AppStream 2.0
ssh
Amazon S3 (Data Egress Bucket)
TRE Data Lake Account
Amazon S3 (Datasets)

*Figure 2. Schematic of Secure Enclave capability within the TREEHOOSE architecture.*

The working model in the PoC is shown in Figure 3 as a three-step process. The TRE-compatible Secure Enclave can now be shared with a third party who wants to ingress their algorithm within the TRE (Step 1 in Figure 3). The third party includes their algorithm within the secure enclave image and encrypts it using their own encryption key. The cryptographically secure image can be shared with the TRE administrators safe in the knowledge it cannot be exposed without the use of the third party's key (Step 2). An approved user of the TRE can run the algorithm within the Secure Enclave only if they have the valid keys which are approved by the third party via the Key (or encryption) Management System (Step 3). The algorithm is decrypted at runtime, processes the data available at a pre-defined location (either a filesystem path or database), generates unencrypted results files and then re-encrypts when the run is finished (Step 4). During runtime the contents of the enclave are hidden from the host environment. Finally, the enclave container can be deleted by the TRE administrators when it is no longer required. Alternatively, the third party can revoke their encryption key thereby rendering the enclave non-functional.

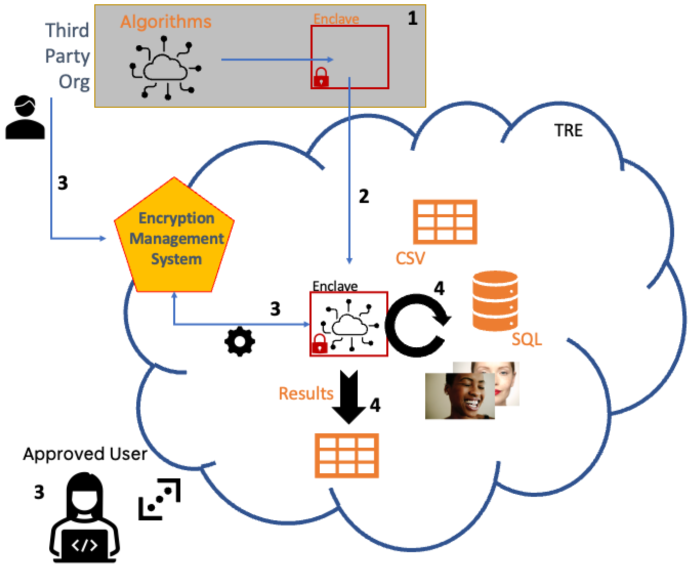

*Figure 3. Secure Enclave Proof-of-Concept.*

It is critically important to note that the security of the TRE and the data within it are not compromised by a secure enclave. The enclave within the TRE is uni-directional and no data from within the enclave is able to leave the TRE. At no point is the enclave allowed to leave the TRE or communicate with the third party. The outputs generated during the running of the enclave's code remain within the TRE and can be used by the researcher in their analyses. An egress request can also be made and must pass the same standard of approval as for any other data egress request.

We have tested the functionality of the PoC with the PICTURES algorithm and it works as expected. However, it is not yet included within the open-source codebase of the TREEHOOSE TRE. We plan to complete an Application Security Review with AWS as part of the next phase of the project which will allow the release of the TRE-compatible secure enclave as open source. As this is still a PoC model, any real-world use will need to go through governance procedures to ensure acceptability and that public benefit is maintained. It is important that TREs maintain their trustworthiness with data owners, the public and other stakeholders.

## 2.3. Work Package 3 – Public and Patient Involvement and Engagement (PPIE)

Two online workshops were held on 14th May and 23rd June 2022 with ~10 participants in each to discuss the use of health data with TREs, explain the current methods used and then listen to the public participants. The aim was to see what their current understanding was and learn how to improve information sharing. The workshop was run with a short initial presentation of the current process and then a series of open questions for comment and discussion. Participation recruitment was from across the UK with members coming from England, Northern Ireland and Scotland. The workshops were led exclusively by the PPIE team members, Antony and Jill, with the project researchers only on hand to either share information or answer direct questions. Discussions were prompted by open questions and left to go in whatever direction the participants wanted.

The workshops' structure was to firstly identify topics and themes of interest by the public in the first workshop and to reflect those in a report which was prepared and shared ahead of the second workshop. The second workshop was used to identify difficult terminology and co-develop language around those terms.

The final report, "Trusted Research in the Public Eye: A toolkit for public engagement for trusted research environments (TREs)", is available here: https://doi.org/10.5281/zenodo.7044482 and a summary is provided below.

### 2.3.1. Summary of PPIE Report

The public workshops were not formal focus groups or research participants meaning that individual quotes or comments cannot be shared, however, a general commentary was provided to summarise the topics and sentiment shared at the events. Five areas were focused upon.

**Information sources**. The public are not very aware of the health data research happening with their data and would value more information regarding how that occurs. There was no clear consensus on where and how that should be done suggesting that different methods should be used to target different members of the public.

**Knowledge gathering/dissemination**. We identified many gaps in the participants' knowledge and understanding of health data research and how TREs are used. More work on bridging the gaps in knowledge needs to be done.

**Patient access to their own data**. Although not a planned topic of discussion, this became an interesting area which highlighted the great variety of access to individual's information across the four nations of the UK. More fine-grained opt-outs would also be of interest to individuals.

**Security**. Security was found to be important in terms of data access and management. Different opinions were voiced regarding cloud vs NHS infrastructure.

**Commercialisation**. There were some strong views around NHS privatisation and how that affected data access and selling data access to the private sector was not considered acceptable.

A glossary of terms with suggested wording for TREs to use when informing the wider public was created and a top-level recommendation was made regarding better dissemination of information regarding TRE and patient data, plus how the wider use of patient data is perceived by the public.

## 2.4. Additional Outputs

- TREEHOOSE presentation at the UKRI Cloud Workshop, March 2022: https://cloud.ac.uk/2022/02/27/programme-for-ukri-cloud-workshop-2022/
- Health Data in Research presentation and introduction to TREEHOOSE slides: https://doi.org/10.5281/zenodo.7040976
- TREEHOOSE presentation at RSECon 2022 Satellite event on Trusted Research Environments, 5th September 2022
- TREEHOOSE poster presentation at Royal Statistical Society annual conference, 12-14 September 2022.

# 3. Impact

Given the sprint only lasted eight months there has been limited capacity to develop concrete impact. However, we have been targeting potential future users and collaborators of TREEHOOSE for the next phases of the project. The main successes are highlighted below:

- Invited to join the Smart Manufacturing Data Innovation Hub (SMDIH) collaboration led by the University of Ulster to provide the TREEHOOSE TRE as an environment for **SMEs to access scalable infrastructure** and specialist digital tools within a secure data environment. This is a **3-year £50 million** project and shows the general applicability of the TREEHOOSE TRE to industrial research outside of health
- Initiating collaborative discussions with other institutions interested in Cloud TRE architecture:
  - Primarily health or health-related data
  - Application of TREEHOOSE to non-health areas such as **environmental data**
  - All in favour of building a community or network to develop a pan-UK AWS interest group
  - In discussion with the Alan Turing Institute and their Azure TRE around **finding commonalities and avoiding silos**

# 4. Next steps

Through the Sprint project it was clear that there was going to be a need to do further work in developing the TREEHOOSE framework. The needs broke down into three areas:

- Broaden adoption to more interested parties (e.g. research institutions, commercial enterprise)
- Develop additional features (e.g. federation, statistical disclosure controls)
- Get wider engagement with the public and patients around health data sharing and trusted research environments

## 4.1. Broaden Adoption

HIC are only one of many Safe Haven organisations or other institutions running a TRE or wishing to run a TRE. As we have found in work leading up to the TREEHOOSE project, adoption of newer technologies is trailing behind the needs of researchers and analysts, meaning there are many existing organisations who are aware they need to embrace change.

We have contacted or been contacted by several UK Universities who have a need or interest in TREs and wish to work with us on a future collaboration to support their adoption of a reference architecture TRE. In the next phase of the project, we wish to formalise these relationships and get funding to support the deployment of TREEHOOSE to very different environments: 1) multi-cloud, 2) analysis of large, historical environmental data, and 3) analysis of commercial/industrial data.

We will build capacity of expertise in developing infrastructure both in terms of knowledge and people across the UK research spectrum, and not only within health.

We are leads on the Alleviate – Advanced Pain Discovery Platform – an HDR UK Data Hub, where one deliverable is to enable TRE usage by data partners for providing secure access to their data. The TREEHOOSE will be a perfect solution for partners who do not already have TRE architecture.

## 4.2. Develop Features

An advantage of having created an open source TRE is that it can be used as a platform to develop features on top of it which will benefit many if not all users. Then when a group wants to use the TREEHOOSE TRE they can get the additional features with little extra effort.

HIC were one of the leads on the CO-CONNECT project which has enabled a federated architecture for cohort discovery and laid the groundwork for federated data sharing and analytics. This will be a fundamental capability to encourage and support portable and reproducible science within TREs. We are also in discussions with third parties to implement their solutions.

Currently, genomic data is not commonly available nor analysed in TREs. The scalability of a cloud platform means support for HPC workflows can be added to TREEHOOSE, either via AWS Batch, or the more portable and open source Slurm workload manager. For example, we will be able to integrate existing bioinformatics NextFlow workflows almost transparently.

HIC led another DARE UK Sprint project, GRAIMatter, which identified specific needs for statistical disclosure control required with AI/ML studies. The project made recommendations to support data governors responsible for managing TRE data egress, including examples of additional “SafeModels” tooling. We can build this tooling into the egress interface of the open source TRE.

Finally, one weakness of cloud TRE solutions is the potential risk of vendor lock-in. With the release of the ATI TRE on Azure there is an opportunity to build common or agnostic toolsets to provide diversity and aid adoption of cloud TREs without silos.

## 4.3. Engagement with Public and Patient Groups

As part of all the DARE UK Sprints it has been clear how much positive impact PPIE members make to the projects and the DARE UK initiative overall. We have seen that the general view of the public and patients is very positive in regards to the use of healthcare data for public benefit. However, more needs to be done. Groups not usually represented and hard to reach such as the young, vulnerable, minority or stigmatised groups also need their voices heard. These groups may have particular concerns regarding their data and how research may impact them. The current outbreak of Monkeypox demonstrates that not all groups are affected equally.

Transparency is a fundamental aspect of GDPR and other data privacy laws so we need to work to be more transparent regarding how sensitive data are used, described and presented in research. The PPIE groups have an important role in developing this work.

## 5. Acronyms

AWS – Amazon web services. A supplier of cloud computing infrastructure.

DSA – Data sharing agreement.

GDPR – General Data Protection Regulation. The EU-developed data privacy and protections laws.

HIC – Health Informatics Centre. Based at the University of Dundee, operates a Scottish regional Safe Haven

IP – Intellectual property.

ML – Machine learning. A computation process for learning statistical patterns from training data to apply on other examples of the same type of data.

SWB – Service Workbench. An AWS application for managing cloud-based projects.

TRE – Trusted Research Environment. A secure computational environment that complies with the “5 safes” for analysis and research on sensitive data.

## 6. Acknowledgement

This work was funded by UK Research & Innovation, Grant Number MC_PC_21032, 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 ADR UK (Administrative Data Research UK).

## 7. Authors

Dr Christian Cole – University of Dundee - PI of the TREEHOOSE project, on behalf of the TREEHOOSE team:

Dr Simon Li – University of Dundee

Dr Chaung Gao – University of Dundee

Dr James Sutherland – University of Dundee

Jillian Beggs – Public participant

Antony Chuter – Public participant

**DOI: 10.5281/zenodo.7085505**

**November 2022**