

Our Future Health

Trusted Research Environment

Procurement - Requirements

Version: 3 March 2022

Contents

1	Background and context.....	3
1.1	Accreditation	3
1.2	Accessing Our Future Health as a researcher.....	3
1.2.1	Public data browser	3
1.2.2	Access Process.....	4
2	Description of TRE requirements.....	4
2.1	User interface	4
2.1.1	User groups	4
2.1.2	Usability.....	5
2.1.3	Web portal access	5
2.2	Separation of workspaces	5
2.3	Cohort Browser	6
2.3.1	Genetic data browser	6
2.4	Analytic tools.....	7
2.4.1	Interactive notebooks.....	7
2.4.2	Statistical tools.....	7
2.4.3	Machine learning.....	7
2.4.4	Genetic tools and pipelines	7
2.4.5	Command line interface	8
2.4.6	Building workflows.....	8
2.4.7	Encouraging collaboration and reproducibility.....	8
2.4.8	Scalable distributed workflows.....	8
2.4.9	Job monitoring.....	8
2.5	Access to compute	9
2.6	Airlock.....	9
2.6.1	Import.....	9
2.6.2	Export	10
2.7	Hosting.....	10
2.7.1	Cloud Infrastructure.....	10
2.7.2	Data centres.....	10
2.8	Data governance	10
2.8.1	Data controllership	10
2.8.2	Data management.....	10
2.8.3	Participant withdrawal	11
2.8.4	De-identification	11
2.8.5	Authentication and authorisation	11
2.8.6	Auditability	11
2.9	Billing.....	11
2.10	Support.....	12

2.10.1	Researcher TRE support.....	12
2.10.2	Documentation	12
3	Future-proofing and portability	13

1 Background and context

1.1 Accreditation

The default route for the majority of researchers accessing the Our Future Health resource will be through the Trusted Research Environment (TRE) hosted by Our Future Health that we are procuring (as Lot 2) through this process.

Our Future Health will establish and maintain an accreditation process that will likely be run by a third party. The process will cover aspects such as access controls, security of the environment and control of the boundary, auditability, protection against re-identification and compliance with industry-standard security and information governance standards and controls. It will be the responsibility of Our Future Health to ensure that our TRE environment passes the accreditation process, and the supplier will be required to cooperate with this process in providing security and data assurances based on those listed in this document.

1.2 Accessing Our Future Health as a researcher

1.2.1 Public data browser

Please note, the public data browser is considered **optional** and will not be evaluated in this bid.

The public data browser will be an open resource that will showcase aggregated, summary level data from the cohort. The public data browser will be accessed via our public researcher facing website (not within the TRE). This tool will be primarily used by researchers to enable:

- discovery of the dataset, exploring data available to them within Our Future Health.
- preliminary hypothesis generation through visual queries of the dataset.
- feasibility of research, estimates of sample power and supporting grant applications

The browser will be expected to support both phenotypic and genotypic data browsing.

Researchers may wish to explore and query:

1.2.1.1 Phenotypic datasets

- Participant demographics (e.g., questionnaire)
 - Number of participants that have answered the question across the cohort as percentage
 - Relevant visual graphs of response distribution
 - Breakdown by age at recruitment, sex at birth, ethnicity
- Disease profile of the cohort (e.g., from fields in NHS linked datasets)
 - Disease counts

1.2.1.2 Genetic datasets

This type of browsing would allow researchers to interrogate the cohort genomic datasets on an individual SNP level. This would inform researchers whether their genetic study is reasonably well powered for robust results. Researchers will be allowed to browse genetic data by querying:

- rsID (e.g., rs12108521)

- SNP chip ID
- Chromosome (e.g., Chr 4)
- Base location (e.g., 99786262-99786262)

Researchers shall have the ability to save and export the selected cohort definition to facilitate submission of research proposals through the access process.

The public data browser will take measures to ensure there is no risk of re-identification of participants, such as limiting the amount of cross tabulation of data and obscuring counts when below a specified range.

1.2.2 Access Process

Please note, software to manage the Access Process is considered **optional** and will not be evaluated in this bid.

Registered Researchers shall submit comprehensive research proposals (including the scope of data requested) to the Access Board for review. Researchers will be required to select groups of variables that relate to the context of their Approved Study (e.g. health-related questionnaire data, HES data). Researchers may be asked to specify a cohort of interest; however, it is possible that data of all participants will be made available to Researchers. The Access Process is currently under development and will be made available in 2022.

It is expected that the Access Process will be supported and facilitated through software within the researcher platform. Researchers shall be able to submit a study for approval via the portal, track progress and manage the study following decision of the Access Board. The platform may also support the Access Board members in reviewing and collaborating in the decision process.

2 Description of TRE requirements

This section outlines Our Future Health's expectations of the technical requirements of the TRE.

2.1 User interface

2.1.1 User groups

Our Future Health will be the UK's largest ever health research programme. As such, we expect to provide a platform for a diverse variety of researchers. These users will have differing levels of technical ability, diverse research interest and requirements for analysing datasets.

As the platform will be used by a wide range of researchers, from diverse fields and levels of expertise, it is important that the user interface follows industry good practice in design usability. We envisage the following design principles to guide Suppliers:

- Make it easy to be data-led. Those already comfortable with data can flex their skill, while those less comfortable are provided with what they need to make it meaningful.
- Grow confidence at every step. Every interaction with the data, from browsing to analysis, should build on existing knowledge and make people feel like power users over time.
- Re-use brain power. Where possible, allow people to build on work previously done, whether it's cohort creation, quality control or pipelines.

- Effort where it matters. Seamless support, documentation and permissions management so researchers can focus on the problems that matter rather than administration and raising tickets.
- One-team working. Team members can work on what's relevant for them, share what's critical for teammates and external collaborators and contribute to wider community in-platform.

In the early stages of Our Future Health, we expect to focus on the requirements of researchers from either academic, charity or industry settings, who are familiar with working in TREs and have experience with large public datasets with similar composition. These researchers may be epidemiologists, machine learning researchers and computational biologists. Suppliers shall be able to demonstrate how their platform caters to these users.

2.1.2 Usability

Suppliers shall demonstrate that they consider usability an underlying principle within their software development lifecycle, addressing usability requirements through user-centred design and undertaking usability testing as a regular part of the test cycle.

Over time we expect that the needs and user groups will evolve and we expect the Supplier to work with Our Future Health to ensure continued product improvement through user analysis.

2.1.3 Web portal access

Registered researchers shall be able to access the TRE via a web portal, provided by Our Future Health, (from multiple commonly used browsers) and through an API providing programmatic access, subject in both cases to appropriate user authentication and authorisation.

The TRE shall integrate with Our Future Health's API to confirm a user's status as a registered researcher. The API will use common industry standards such as, but not limited to, OAuth.

The user interface shall be accessible, and support standards such as, but not limited to, the Web Content Accessibility Guidelines (WCAG). For our participant-facing services we are designing to an AA level.

The TRE branding shall be configurable to reflect Our Future Health branding.

2.2 Separation of workspaces

The TRE shall provide a separate cloud workspace for every Approved Study, where researchers can conduct their research. The workspace shall be isolated from the rest of the TRE, such that actions within an individual workspace (malicious or otherwise) solely impact that workspace.

The TRE shall integrate with Our Future Health platforms to ensure that Registered Researchers shall only have access to approved data relevant to their Approved Study, and only have access to the relevant workspace(s).

Enabling collaboration through separation of repositories across multiple approved studies is a core functionality requirement of the TRE:

- A researcher could participate in multiple approved studies.
- Multiple researchers may be active within the same Approved Study, and we expect that different researchers may be working on distinct aims within an Approved Study.
- Individual researchers may be collaborating on multiple approved studies. However, data shall not be transferred between Approved Studies.

The account holder, who is the lead applicant of the study, shall be able to modify workspace permissions and invite other Registered Researchers associated with the Approved Study. The account holder will also be responsible for paying all the cloud activities in the workspace, which includes incurring cloud costs of collaborators within the workspace. Please refer to Lot 3 Researcher Billing Requirements for further details of this requirement.

Administrative controls shall be available to Our Future Health staff to modify workspace permissions.

2.3 Cohort Browser

Researchers shall be able to explore the available data in the Approved Study to define one or more cohorts per project before they start their analysis. Researchers shall be able to apply relevant inclusion and exclusion criteria, and filter by one or more data fields of any attribute or data element as a criterion.

It is expected that users shall be able to visualise their defined cohort via interactive graphs and charts to inspect attributes such as data distribution and counts.

Researchers shall have the ability to save and export the defined cohort to enable repeatable analysis.

2.3.1 Genetic data browser

Researchers who have access to genetic data may need to create a subset in order to streamline their analysis. This shall include filtering the genetic data for regions of interest (examples include, but are not limited to, CNV & SV), SNP/point mutation, genes and SNP chip IDs. There shall be an expectation to view a list of variants with associated annotations, including, but not limited to, variant type, impact/consequences of variant and allele frequencies (both within the entire dataset & public). It is expected that researchers shall be able to filter based on these attributes.

Variant annotations from popular databases and repositories shall be supplied by Our Future Health. However, researchers shall be able import their custom variant annotation datasets for visualisation and filtering if they wish to do so, which shall be subject to import restrictions described below.

Researchers shall be able to use genetic visualisations to help digest the list of variants. Examples of such visualisation include:

- displaying allele frequencies (lollipop graphs) in the cohort,
- displaying a genome browser to allow for more dynamic and finer browsing of the chromosomal regions to aid identification of structural variants and SNP across the cohort mapping to other publicly available databases (examples include, but are not limited to, ClinVar, gnomAD, Decipher).

2.4 Analytic tools

Commonly applied tools and pipelines to support research and analysis of complex data will be required.

2.4.1 Interactive notebooks

Researchers shall be able to develop code using an interactive notebook (such as a Jupyter Notebook) within the TRE to allow for bespoke, complex querying and analysis of the data. The notebook shall support prevalent tools and packages used for statistical analysis such as, but not limited to, Python, R or Julia.

Researchers shall be able to collaborate with other researchers in their Approved Study and generate publication-ready graphs and tables.

Support shall be included for notebooks that require access to distributed processing (examples include, but are not limited to, Apache Spark) for the running of specialised tools, such as HAIL2.

2.4.2 Statistical tools

We would expect that prevalent statistical packages would be available to researchers for analyses. Such tools may include, but not limited to: SAS, Stata, SPSS and R Studio. We are aware that there may be complexities around licensing of these tools and will work together with suppliers to find solutions we can offer to serve a wide range of the research community.

2.4.3 Machine learning

Researchers shall be able to train and evaluate machine learning models through access to appropriate repositories and tools (such as PyTorch and TensorFlow) and workflow, orchestration and pipeline MLOps tools commonly used for machine learning workloads.

2.4.4 Genetic tools and pipelines

The TRE shall support a broad repository of commonly used tools/pipelines for batch level genetic data which shall be easy to search. All tools and pipelines that are made available to researchers shall have version control to enable reproducible analysis. Researchers shall be able to select a tool or libraries and specify a version (e.g., Hail Version 0.2.81 via Jupyter Notebook).

Genomic data handling tools shall include:

- Processing (e.g., GATK),
- Harmonisation,
- Manipulation (e.g., VCFtools)
- Annotation (e.g., SnpEff, VEP).

2.4.4.1 Genome and phenome wide based analyses

A common use case expected is the study of association between the cohort's genetic make-up against phenotypic data points. Researchers shall be able to conduct basic association-based analyses (for example GWAS) using their defined cohort.

Researchers shall be able to use this output as well as other downstream analysis, such as generating genetic risk scores (e.g., polygenic risk scores), and more phenotype focused analysis (e.g., PheWAS).

The TRE shall enable the provision of interactive applications, for example, tools appropriate for filtering of genetic data and visualisation tools such as GWAS and/or PheWAS Manhattan plots.

Genetic analysis tools shall include, but not be limited to:

- Association analysis (e.g., PLINK, SAIGE, Bolt LMM, METAL),
- Genetic risk score (e.g., PRSice-2),
- Relatedness calculator (e.g., GRM),
- Colocalisation (e.g., coloc).

2.4.5 Command line interface

Some projects may require more complex querying and analysis that is arduous via a web-based UI, but easier via a command line interface (CLI).

It is expected that researchers shall be able to easily run all the features and functionality (e.g., create/manage projects, analysis of data) via a CLI.

2.4.6 Building workflows

Researchers shall be able to create custom workflows using existing tools and pipelines on the TRE platform or from those that are imported. The workflow shall be simple to build and straightforward to modify and edit. Researchers shall have the ability to create workflows via the web browser or via CLI.

2.4.7 Encouraging collaboration and reproducibility

We would like to encourage researchers to collaborate both within and between workspaces. We are interested in capabilities that allow researchers to publish code, notebooks or workflows such that other researchers can re-use or build upon their work, or contribute to it collaboratively.

2.4.8 Scalable distributed workflows

We will need to provide a system that facilitates complex workflows, providing scalability via distribution across multiple compute instances. Examples of this include Nextflow, Snakemake, Cromwell/WDL. This method shall provide options to control the scale and cost of a job, e.g. VM shape and number, job duration, priority, etc.

2.4.9 Job monitoring

Researchers shall be able monitor jobs that are running in their workspace via both web UI and CLI. The status of jobs shall include whether the job has started, running, completed or failed. A historical log of previous jobs shall be accessible to researchers. Furthermore, researchers expect to be able to manage longer running jobs by selecting the preferred compute resource pricing model (e.g., spot, reserved, or on-demand).

Researchers shall be able to view the costs of a job, with the account holder being able to terminate jobs that are costly. Please refer to Lot 3 Researcher Billing Requirements for further details of this requirement.

2.5 Access to compute

We expect there to be a wide range of research carried out on the data, some of which – such as training machine learning models - may be computationally heavy. The TRE shall have the plasticity to enable researchers to efficiently fit the platform to their own needs, without expending more resources than necessary.

The Supplier shall make available a range of compute instances that can be selected by the Researcher based on their analysis requirements. Researchers shall be able to use a range of instances, available out of the box, with configurations to include additional resources as required. The Supplier shall be able to provide guidance to the Researcher around available and recommended instances, including:

- general purpose analysis
- compute-intensive analysis
- memory-intensive analysis
- specialised compute instances (including GPUs).

2.6 Airlock

The TRE shall allow researchers to import their own data, software and tools. Researchers shall be able to export aggregate results, models, code. However, export of participant-level data will not be permitted. Data transfer shall be accomplished using industry standard tools (e.g. SFTP).

Safe outputs shall be ensured by an Airlock process, by which export of screened and approved information that is non-disclosive can be permitted following a review of the request. During the first six months of operation, we expect this airlock process will be handled by an Airlock Manager role via manual review. The Airlock Manager will be an employee of Our Future Health. The Supplier shall provide the framework to enable the Airlock Manager to review and approve requests from researchers. Researchers shall be able to initiate and receive a response regarding an export request in a timely manner. Suppliers shall be able to provide evidence on usability and key performance indicators, such as time to decision for this process.

With the number of projects and researchers increasing over time there is an expectation the manual airlock process would move towards a more scalable model and we would invite Suppliers to comment on their product roadmap relating to this feature.

2.6.1 Import

Importing of tools and data to a workspace shall be controlled. Researchers shall be able to safely upload code, container images (e.g. Docker), new tools or to make use of additional libraries.

Researchers shall have the opportunity to import their own datasets on to the TRE environment to conduct meta or cross analysis.

2.6.2 Export

Researchers shall be able to download and export summary level results from their Approved Study for further analysis using their own infrastructure.

Researchers shall be able to export or synchronise code developed in the TRE. Our Future Health will be developing policies to ascertain whether this can happen directly via source control systems (e.g., GitHub, GitLab) or via an airlock process.

Researchers may want to share preliminary findings or provide limited read-only access of results to close collaborators. Multiple export requests may be potentially burdensome to users and could also create extra security risk through cumulative requests. This process may also be appropriate to enable peer review for publication of research.

2.7 Hosting

Please refer to Lot 1: Cloud Requirements for further details on cloud infrastructure requirements.

2.7.1 Cloud Infrastructure

The TRE shall be hosted within the same public cloud used for all Our Future Health services.

We envisage the TRE Supplier deploying their platform within our cloud account and continuing to maintain it as a Software-as-a-Service there, using appropriate separation from other Our Future Health services. We will negotiate the details of this deployment during the integration between winning suppliers of the three lots.

2.7.2 Data centres

The primary data store and the TRE shall be hosted within the UK such that participant data does not leave the UK.

2.8 Data governance

2.8.1 Data controllership

Our Future Health is the data controller of the primary participant data. Within a TRE workspace for an Approved Study, Our Future Health and the researchers shall both be independent data controllers of the data used for the study. The TRE Supplier shall act as a data processor.

2.8.2 Data management

The Supplier shall support Our Future Health's data management principles from the outset, to ensure that Registered Researchers can only interact with Our Future Health data in the context of their Approved Study. The data made available by Our Future Health for research will not be static, with future data releases including additional participants, updated data and new data types. Our Future Health will be receiving new data continuously. However, it is expected that data made available to researchers will be refreshed at specific intervals, for example quarterly.

Researchers shall be able to store intermediate data tables or files, results, models, code in their workspaces, and have a reasonable expectation that this is secure and backed up and available for the lifetime of their Approved Study. There may be researchers who wish to store a subset of the

cohort data within their workspace, for example to guarantee a specific version that does not change to ensure their analysis is reproducible.

2.8.3 Participant withdrawal

Participants can choose to withdraw from the programme at any time. The datasets provisioned by Our Future Health to be made available within the TRE shall take account of any participant withdrawals. The supplier shall not be expected to process withdrawal requests.

2.8.4 De-identification

Researchers will be provided data that has been de-identified, with a fresh set of research participant IDs for every approved additional study.

Please note, the process for de-identification is considered **optional** and shall not be evaluated in this bid (it may be provided by Our Future Health via a bespoke system, it may be purchased separately, or it may be supplied by the winner of this procurement Lot).

2.8.5 Authentication and authorisation

The Supplier shall include appropriate levels of authentication and authorisation controls to prevent unauthorised access to systems and application, so that only registered researchers are able to access the workspaces and datasets approved for their approved studies. The Supplier shall employ processes and systems and impose user credential requirements to minimise unauthorised user access. As stated in section 2.1.3, Suppliers shall integrate with Our Future Health systems to determine the status of a registered researcher, using a commonly used API such as OAuth. In the future the Supplier shall collaborate with Our Future Health to investigate additional standards such as those from GA4GH for researcher passports.

To further minimise risk of unauthorised access, we expect the Supplier to work with Our Future Health in agreeing user credential requirements, such as two-factor authentication, password requirements, credential lifecycles, and other policies.

2.8.6 Auditability

Researcher activity in the TRE shall be logged and auditable to confirm that researchers have stayed within the boundaries of their Approved Study. The Supplier shall ensure that all aspects of the TRE, researcher access, cloud, data usage and data exports are auditable. Effective logging and monitoring shall be utilised to record events and generate evidence.

2.9 Billing

Researchers shall pay for their own cloud storage and compute when working in the TRE, to allow them to scale up their work and access resources as required. The Supplier of the TRE must be able to work with the winner of Lot 3 (if not the winner themselves).

We expect that the account holder will have the option to set thresholds and caps within the workspace. In the case that no threshold is set by the account holder, account holders should be notified of any activity (e.g., analysis, downloading summary data) on the platform that would be deemed as excessive.

The collaborators within the workspace should have some way of knowing how much allowance they have remaining. Additional features that enable better cloud cost management such as estimating cloud costs prior to running large scale analysis would allow researchers to be aware of what the costs of executing that run would be. Furthermore, collaborators should be provided with estimated egress or ingress costs prior to downloading or uploading results or data.

Please refer to Cross Lot Requirements and Lot 3 Researcher Billing Requirements for further details on these requirements.

2.10 Support

Evidence of operational maturity is an important requirement of the TRE. The Our Future Health technical operations team shall need access to a support function from the Supplier in case of technical issues particularly with our production systems. We shall expect to see:

- 365x24x7 access to third line support.
- A sliding scale of response times, with a maximum of 1 hour in the case of production systems that are down.
- Documented, transparent processes and SLAs on patching frequency, TRE availability, upgrade frequency, CSIRT et al.
- Experience of providing services to a range of organisations (academic, healthcare, industry)
- The ability to scale rapidly when required

We are currently assuming that Our Future Health will run a service desk to handle enquiries from researchers, via email or phone. We shall require the Supplier to offer 3rd line support for issues that relate to the TRE platform.

Communication shall be built into the TRE platform by which researchers may be informed of priority information, such as service updates, planned down time or support notifications within the TRE. Our Future Health may provide copy or content for notifications within the TRE from time to time.

2.10.1 Researcher TRE support

The TRE shall host different user groups, with various levels of expertise. We hope to reduce friction by ensuring all users, irrespective of their technical expertise, receive guidance and support on how to interact with the TRE. Such training and documentation shall include topics such as:

- Importing or adding custom tools (e.g., custom-made software)
- Data import
- Onboarding help
- The command line interface & interactive notebooks
- Tutorials on running pipelines
- Video-based tutorials

2.10.2 Documentation

User research has shown researchers have a strong preference for written documentation on the platform as the main type of support. A service desk (provided by Our Future Health) and online

knowledge base (provided by the TRE Supplier) shall allow users to self-serve as far as possible through onboarding materials and user manuals that shall be accessible across the TRE and may have links from the researcher website, allowing users to quickly access it at any point during their time on the platform.

3 Future-proofing and portability

Our Future Health is a long-term health research programme that may last decades. It is expected that the TRE shall be iterated and improved to maintain the service as an industry leading platform for researchers. We seek commitments the Suppliers shall be equally committed to continual product development as new data, tools, and insights from users emerge.

We must consider that all components of our technology stack shall be replaced over time and therefore flexibility and portability of aspects of the TRE shall be considered.

Longer term, Our Future Health is committed to exploring and trialling techniques that allow efficient data analysis across multiple TREs, thus minimising the need to replicate datasets among TREs. These may include the use of synthetic data and federated machine learning or data science techniques that are the subject of ongoing research across the community.

The standards for accreditation of the TRE will be reviewed and renewed periodically with changing governance and developments in technology.