DPUK Data Portal Service Brochure

Providing Secure Data Access for Dementia Research

Website : portal.dementiasplatform.uk
Enquires : helpdesk@chi.swan.ac.uk

Dementias Platform UK | Data Portal

Welcome to the DPUK Data Portal

Providing Secure Data Access for Dementia Research

DPUK's Data Portal has been developed to enable researchers anywhere in the world to be able to access high quality, multi-modal data from more than 50 population and clinical cohort studies, comprising records from over 3.5 million people.

>85 Multimodal Datasets

>2300 Access Requests

>40 Countries

>560 Project Applications

>1100 Total Researchers

The data portal is free to use and available from anywhere in the world. Over the last 10 years DPUK has built a robust and flexible data management environment that supports multiple data types and has built a growing community to help develop the governance to support innovation.

Table of Contents

Introduction	Page 1
Contents	Page 2
Trusted Research Environment	Pages 3-4
Data Management	Pages 5-6
Data Discovery & Curation	Page 7
Computational Resources	Page 8
Neuroimaging	Page 9
Genomics	Page 10
Federation	Page 11-12
Hubs & Collaborative Spaces	Page 13
Register & Trials Support	Page 14
Artificial Intelligence	Page 15
Synthetic Data	Page 16
Large Language Models	Page 17

Trusted Research Environment (TRE)

Accessing Data within a Secure Environment

DPUK Data Portal is a secure trusted research environment that incorporates the five safes as defined by Health Data Research UK (HDRUK). These measures and controls along with our ISO27001 and DEA Accredited SeRP infrastructure ensures that the data we hold and provide access to, are protected to the highest standards.

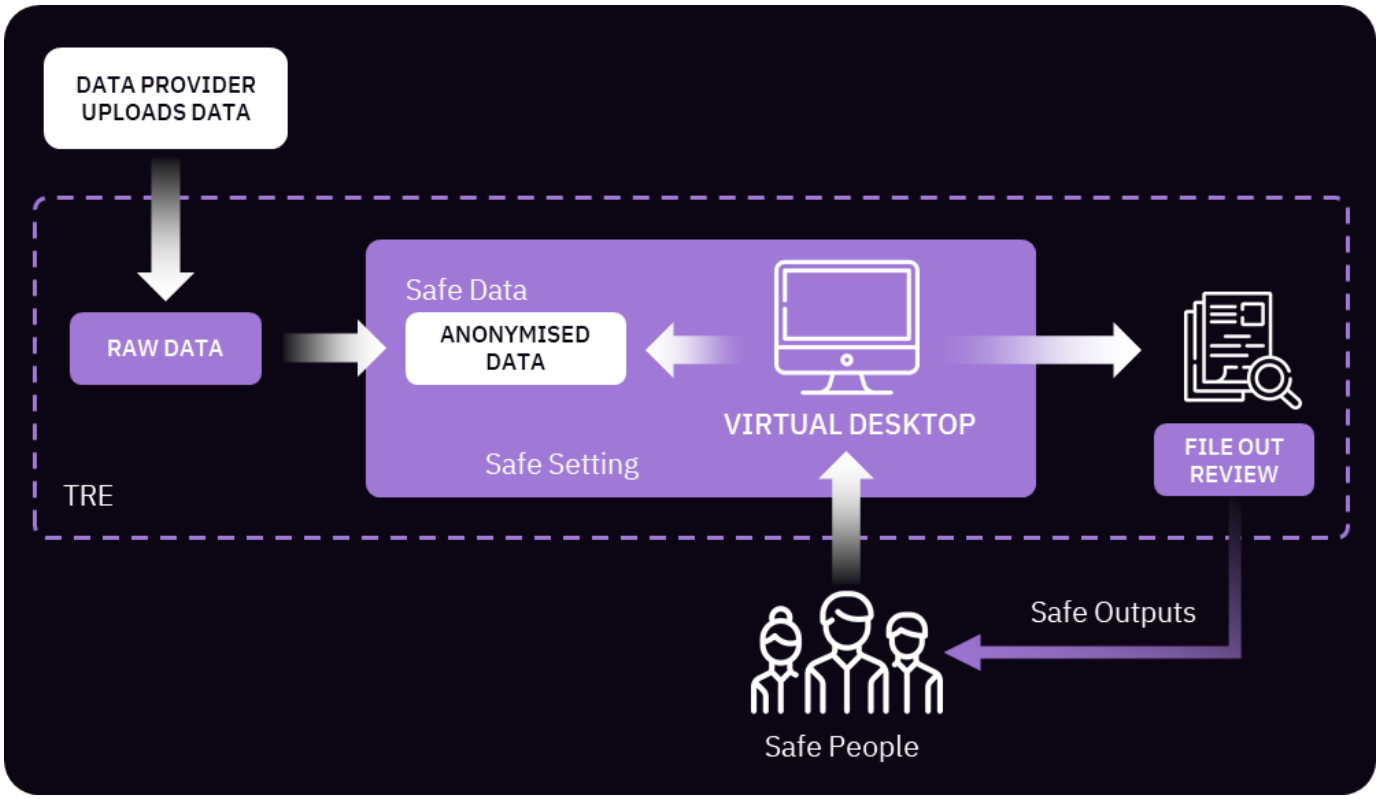
Key Features

Data Never Leaves
Approved researchers access the data via a remote virtual desktop - Data never leaves the secure data environment

Analysis Environment
Researchers access specialised virtual desktops preloaded with analysis software

Data Filtering
Automated variable level provisioning ensures only the data approved is accessed

Output Disclosure Checks
All outputs are checked to ensure that they do not contain any row-level data



Trusted Research Environment (TRE)

The Five Safes

Maintaining the security of participants' identities in the data we manage is fundamental to everything we do. DPUK adheres to the Five Safes framework to ensure data is accessible for research while prioritising its safety and security.

1

Safe People

Researchers must be part of a research organisation and each user signs a Data Access Agreement promising to protect the confidentiality of the data at all times which is countersigned by the researchers organisation.

2

Safe Projects

Researchers submit an application for data access to carry out a specific project. Researchers must demonstrate that the use of the data is appropriate and ethical, and that the research benefits the public.

3

Safe Settings

On approval researchers are given access to data requested via a Virtual machine (VM), where they can analyse the data on our secure systems. Data cannot be removed by the researcher from the secure environment.

4

Safe Data

To further protect the identity of participants we de-identify the data by removing names, addresses and any other details that would directly identify data subjects, before we make them available for analysis.

5

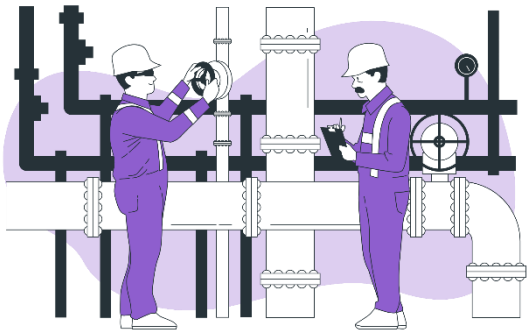
Safe Outputs

As data cannot be removed from the VM, once work has been completed the results can be requested out of the TRE for publication etc. The results are checked by data analysts to ensure they are safe for release.

Data Management

Our Data Management Processes

As a mature data platform we have developed our data management capabilities to be more automated, flexible, auditable, and transparent. Ensuring that we can track the full lifecycle of data within our trusted research environment. The investments into these areas have allowed DPUK Data Portal to extend its data management operations beyond dementia to other Neurological data platforms established within DPUK Data Portal.



Data Ingest Pipelines

Data acquisition pipelines have been developed to ensure that data of any type can be easily and efficiently brought into the portal for onward sharing. Agreements have been adapted and refined to allow multiple use cases. Adaptable tracking workflows have been put in place to ensure that datasets are safely transferred from source, and checked for any quality issues, as well as being formatted for researcher use and stored in our ISO207001 and DEA Accredited SeRP servers.

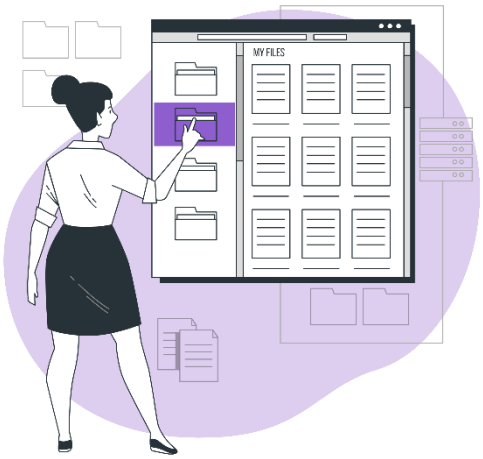
Project Application Processing

Researchers can apply to access datasets using a bespoke application system that enables centralised processing of applications. The system allows applications to be edited in response to feedback from scientific and technical triage staff to provide a more efficient process. The workflow underpinning the system also allows conditional processing so that that different forms and information can be provided automatically to applicants, depending on their specific project needs.

Data Quality Checks

Data and accompanying metadata collected by cohort studies, are stored in many different formats. On ingest, we have developed pipelines for checking the metadata matches the data, that quality is sufficient for researcher access and that there are no potentially disclosive variables, still in the data. The pipeline generates a report and we work closely with the data provider to ensure that the dataset is researcher ready.

Data Management



Data Provisioning

Our ingest pipelines also allow us to tag data and reformat the metadata to a standard template. This allows applicants to select the variables required for their project. The variable selection is then used to automatically provision the data requested using our SeRP automated provisioning. This allows us to provision data quickly with fully auditable logs, whilst simultaneously reducing resource requirements

Output Checks

A key aspect of our Trusted Research Environment is ensuring that data that could be identifiable outside of the security of the TRE are never allowed to leave the environment. To enforce this we manually check all research outputs that are requested to be taken out of the portal, against our robust checklists (check list).

Tracking Project Outcomes

The purpose of the Data Portal is to enable research for the public good. This makes it important for us to be able to track publications, engagement activities, further funding and any other research outcomes. An automated contact collection form has been developed to collect outputs from project leads directly into a database. We also provide each dataset with a Digital object identifier (DOI) so that researchers can credit the appropriate dataset. These can then be directly attributable back to the cohort studies that provided data to the project. The data collected can then be provided to the project leads.

Dashboards

DPUK Data Portal uses Power BI dashboards to surface important data about the project enable dashboards to be created that cover high level overviews (Report), Application and Acquisition tracking dashboards, Discovery tools for researchers and participant selection tools for our clinical trial registers.

Data Discovery & Curation

Enabling Researchers to Discover the Best Datasets for their Projects

To make it easier for researchers to find the data needed for their analyses, DPUK Data Portal have developed a set of discovery tools ranging from high level matrix overviews to in depth variable discovery tools. The key to these discovery tools is the categorisation of variables to a common ontology (c-Surv) developed by the University of Oxford. This bespoke ontology allows variables to be grouped into categories and enables a powerful set of meaningful discovery tools to be developed.

Cohort Acronym	Age at Recruitment	Estimated N	Sex	Start Date	Data access	Administration	Socio demographic	Early life environment	Medical history	Family disease history	Psychological status
ACONF	Birth	12500	MF	1950	Access via DPUK portal	✓	✓	✓	✓	✓	✓
Airwave	18+	53280	MF	2004	Access via DPUK portal	✓	✓	✗	✓	✓	✓
ALSPAC (Children)	Birth	15589	MF	1991	Access via DPUK portal	✓	✓	✓	✓	✓	✓
ALSPAC (Parents)	18+	13, 761	MF	1991	Access via DPUK portal	✓	✓	✓	✓	✓	✓
AMPLE	60+	80	MF	2013	Access via DPUK portal	✓	✓	✗	✓	✗	✓
BDR	65+	2963	MF	2009	Access via DPUK portal	✓	✓	✗	✓	✓	✓
BioFIND	70+	324	MF	2005	Access via DPUK portal	✓	✓	✗	✗	✗	✗
BRACE	40+	2389	MF	1994	Access via DPUK portal	✓	✓	✗	✓	✓	✓
Cam-CAN	18-87	2676	MF	2010	Access via DPUK portal	✓	✓	✓	✓	✓	✓
CamPaIGN	30+	142	MF	2000	Access via DPUK portal	✓	✓	✗	✓	✗	✓

Matrix

Use our Matrix tools to quickly scan and compare data available by viewing which categories of data each cohort contains.

DISCOVER NOW

Dictionary

Our data dictionary provides details of the study data and allows you to explore the variables that has been collected.

DISCOVER NOW

Explorer

An interactive tool that provides visualisations of participant numbers per variable per dataset.

DISCOVER NOW

Computational Resources

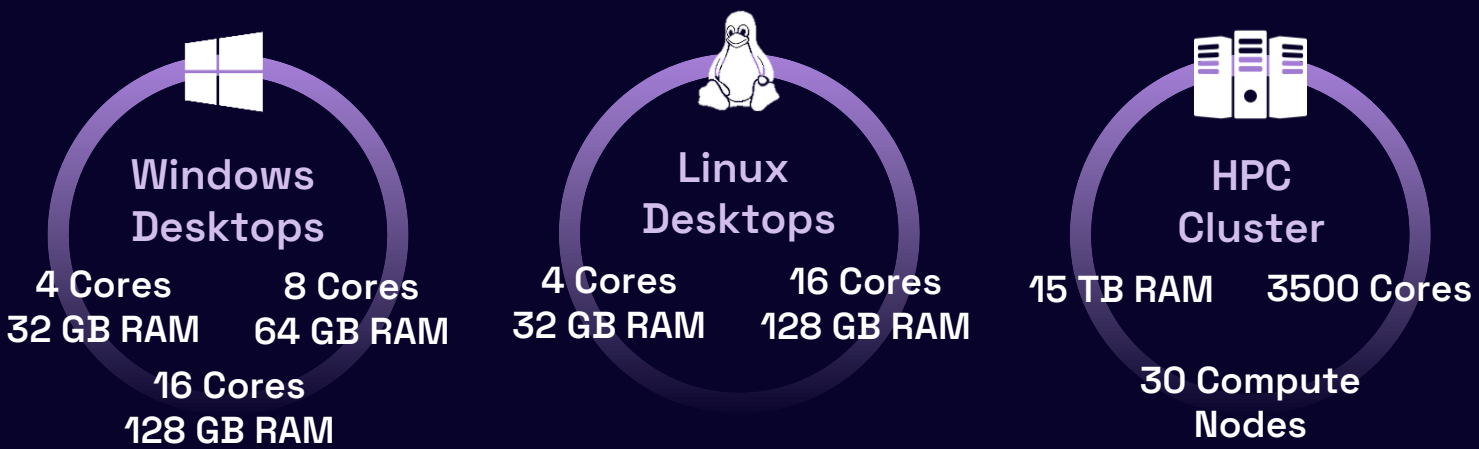
Resources for Advanced Big Data Research

Depending on the requirements of the project, the DPUK Data Portal environment offers researchers many resource options. Both Windows and Linux desktops are available in Standard, Large and Extra Large specifications.

STANDARD		LARGE		EXTRA LARGE	
4 CPUS	32 GB RAM	8 CPUS	64 GB RAM	16 CPUS	128 GB RAM

All of these desktops come pre-loaded with analytical software and specialist tools including: Python, R, MATLAB, Jupyter, Anaconda, SPSS, SAS, STATA, FreeSurfer, FSL and many more. Additionally, you can request new software to be installed within the Data Portal if we don't have it already.

For more advanced projects, we also offer our High Performance Computing (HPC) cluster on a cost-recovery basis. This includes 15TB of RAM, 3500 CPU Cores, GPU, and 30 Compute Nodes.

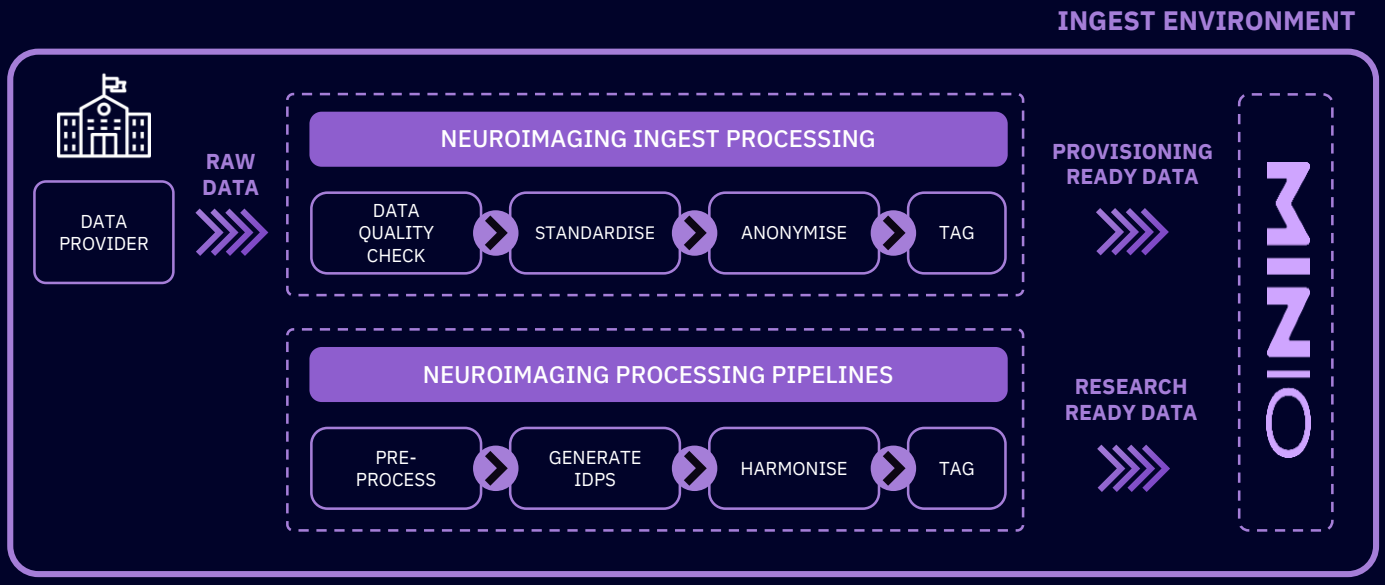


Neuroimaging Data

DISCOVER NOW

Neuroimaging Pipelines & Processes

The DPUK Data Portal brings together over 130,000 brain scans across 20 neuroimaging cohort datasets into a centralised resource, all standardised to the Brain Imaging Data Structure (BIDS). DPUK has worked on developing comprehensive imaging pipelines to anonymise, standardise, process and harmonise data. We have also implemented efficient storage solutions, to enable the provisioning of neuroimaging data to researchers inside the secure environment.



1. Data Upload
Data is uploaded to our environment either through a simple 'drag and drop' SFTP or via the XNAT platform

2. Anonymisation
Structural scans are defaced to remove identifiable features and PII is stripped from DICOM headers and JSON files

3. Standardisation
All imaging data is standardised to the Brain Imaging Data Structure (BIDS). This is done through a semi-automated approach

4. Pre-Processing Pipelines
Pre-processing pipelines implemented in the portal to process scans and provide the derived results

5. Imaging Derived Phenotypes (IDPs)
IDP's are generated from the pre-processing and are stored in CSV files. Anybody else who generates IDP's can also choose to share them

6. Harmonisation
Derived results from the imaging data is harmonised if datasets come from multi-site studies

7. Tagging
All data is tagged with custom metadata to allow filtering of data using permissions and mounting

8. Object Storage Server
Data is stored in our object server which has improved efficiency and lower costs than typical file stores

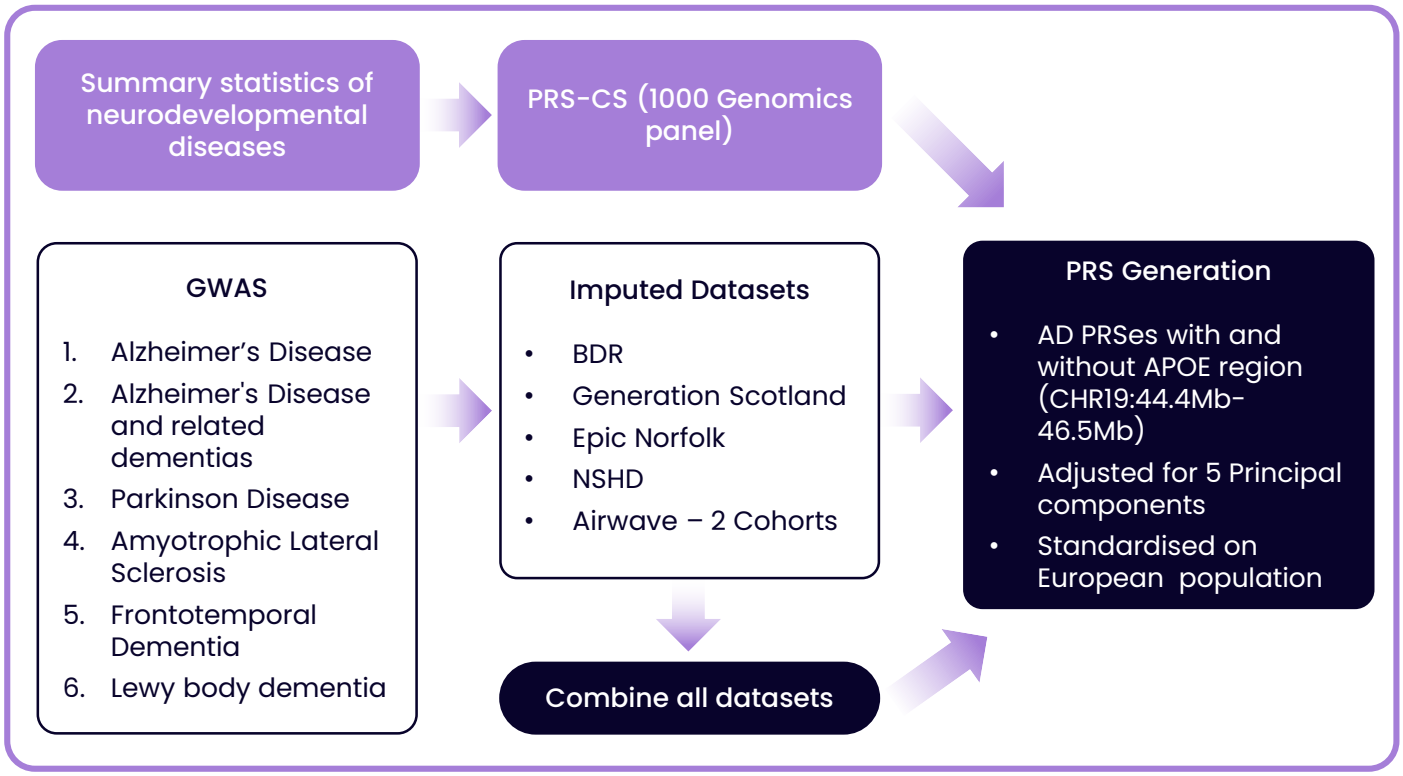
Genomics Data

DISCOVER NOW

Genomics Pipelines & Processes

The DPUK Data Portal brings together raw genomic data and the APOE status of over 100,000 study participants across cohort studies. Given the complexity and heterogeneity of the genetic data, joint analysis is only possible after standardisation and imputation of the data. This is because of genotyping platforms, difference in QC analyses, and the number of overlapping variants, when combined at the individual level.

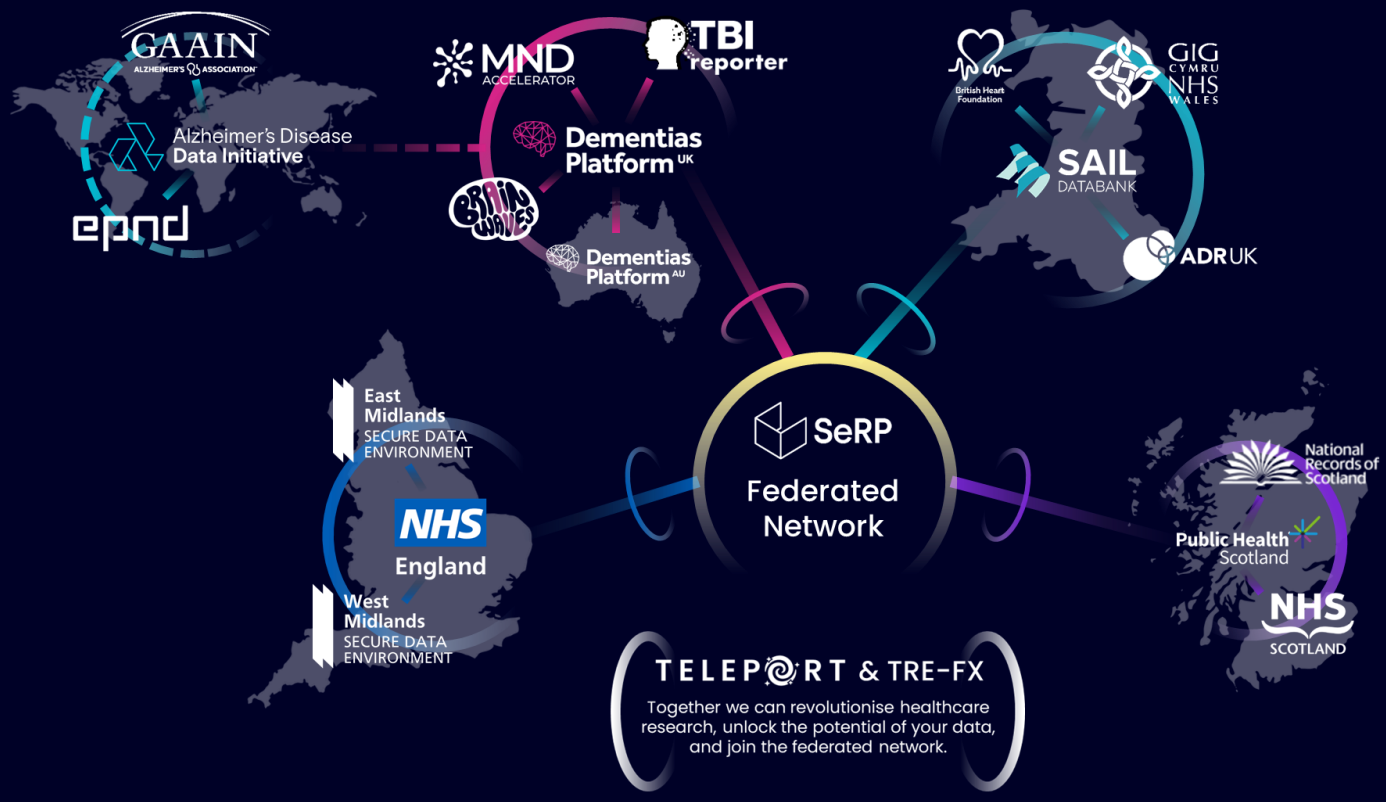
To address these issues, the DPUK Data Portal has developed data processing pipelines with standard QC and data analysis parameters These are open sourced scripts which can easily be adapted by other researchers for their study designs. The pipelines can also be modified to perform other genetic analyses, that is, gene-set /pathway specific Polygenic Risk Score (PRS) calculation with other GWAS summary statistics. These pipelines have also been used to generate PRS scores for multiple datasets within the DPUK data portal thereby enriching the datasets further.

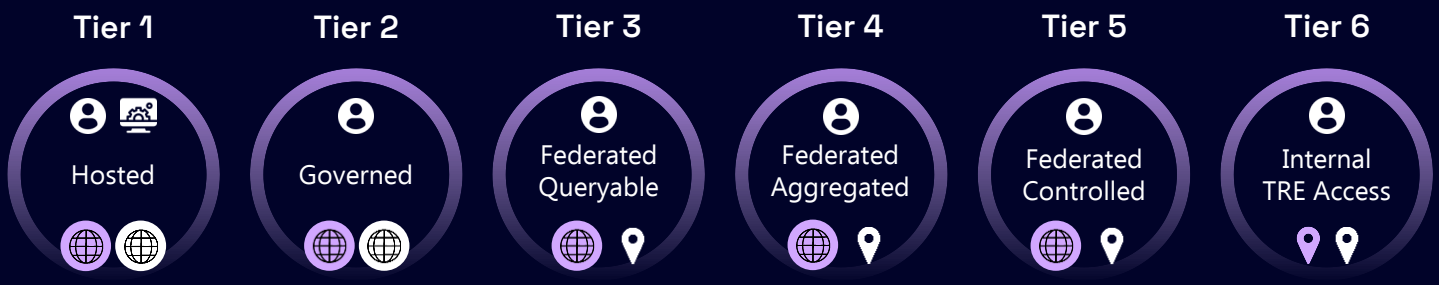


Federation

Opening Up Safe Access to Data

DPUK has developed a unique tiered approach in data governance, that enables researchers to access data across multiple data repositories. This has been achieved via two federated solutions developed by UKSeRP. These solutions enable parallel access to data housed in multiple, physically separated environments (without data moving from their host environments, instead being accessible from its host location to a researcher in a single safe, secure environment).





Human Operated

Machine Automated

Location : Data can be held at any location

Destination : Data can be held at any location

Location : Data is held locally within the platform

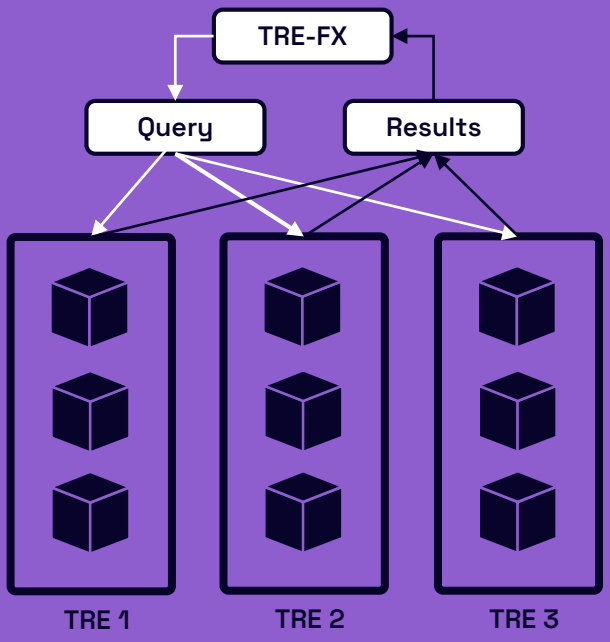
Destination : Data is held locally within the platform

Federation

TRE-FX

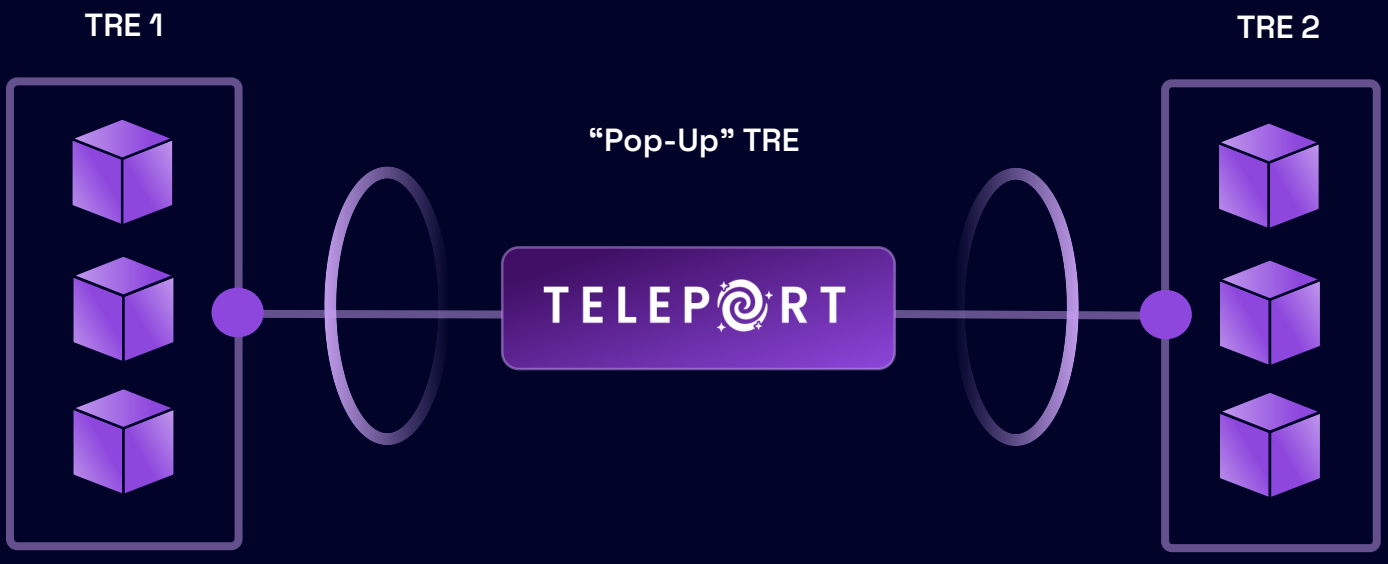
The TRE-FX solution enables packaged analyses, using the same analysis methodology, to be sent to a network of environments in parallel through containerised workloads to draw back results from each environment, with no direct access to the underlying individual-level data other than in viewing its metadata and characteristics through a central workload delivery function.

This technology underpins federated access to DPUK data repositories via the AD Workbench operated by AD Disease Initiative. Through TRE-FX, approved researchers from around the globe can analyse access data held within DPUK along with data held in other data repositories securely.



Teleport

Teleport is a SeRP solution that can be deployed and sits across two or more TREs. The 'pop-up TRE' has full governance control by all participating TREs and researchers wishing to access data across TRE's apply to each TRE and follow their access procedures. Upon approval, researchers are provided access to the pop up TRE through a web interface, which will offer a research environment that is jointly controlled by all participating TREs. This solution provides a very simple user environment allowing the researchers to handle and analyse the data as they would normally within a TRE. However, this solution requires that data sits in an environment outside the original TRE/Data Repository. Nevertheless, that environment is still governed by the data controller.

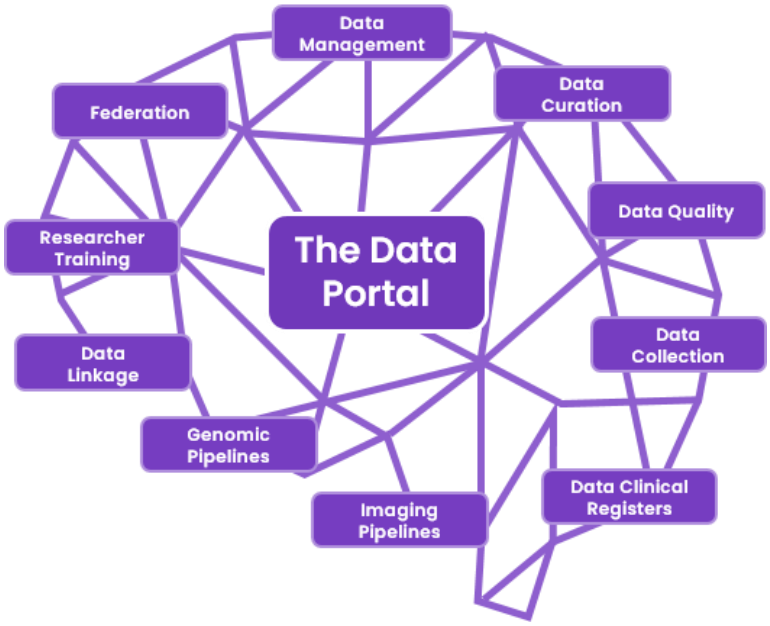


Hubs & Collaborative Spaces

Hubs

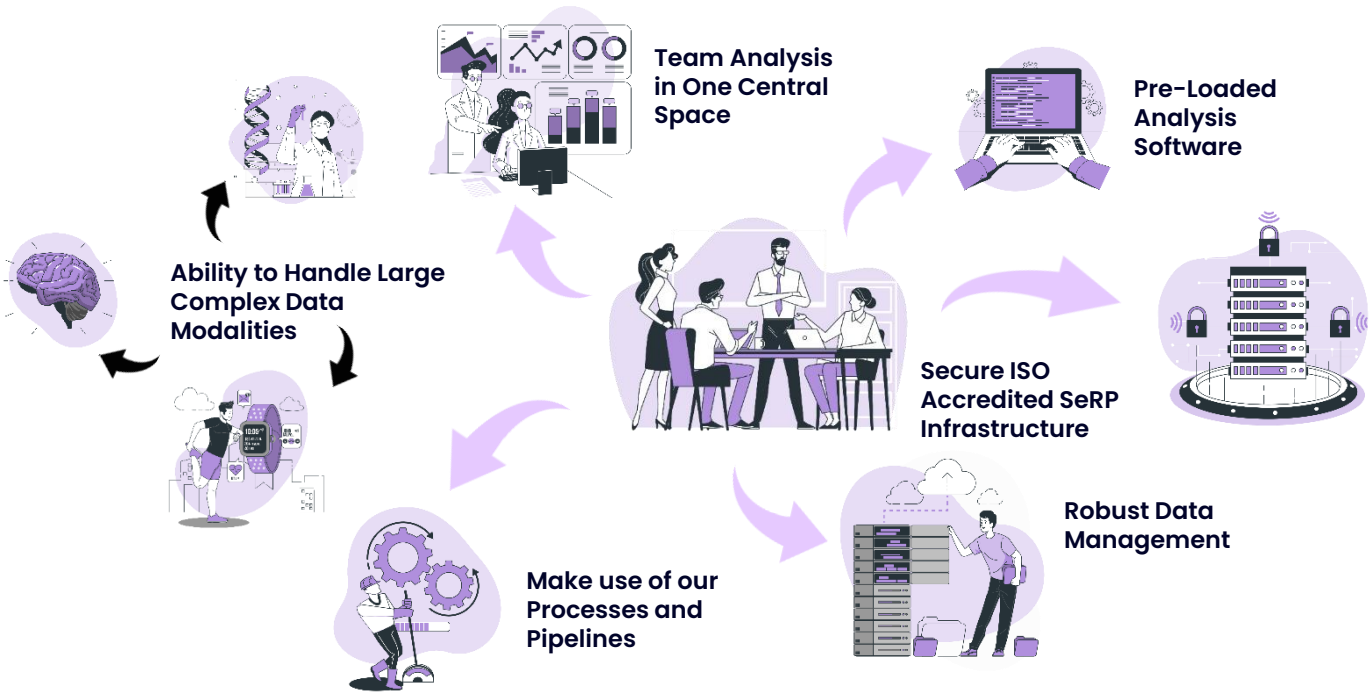
As a mature trusted research environment (TRE), DPUK has established start-up TREs, utilising its already established SeRPUK infrastructure.

At DPUK this is achieved by providing each hub with its own identity (branding), web-presence, and hub-specific utilities, whilst making available the governance, data processing and data analysis pipelines that have already been developed in DPUK.



Collaborative Spaces

Collaborative spaces allow projects to access their own data inside a TRE for analysis. They have their own space within the TRE where their collaborative partners can be provided access to analyse the data in one place. This allows the project to have a solid data management plan and at the end of the collaborative project the data is deposited to DPUK for onward sharing.



Register & Trials Support

DPUK's Pre-Clinical & Clinical Registers

DPUK's pre-clinical registries give researchers access to more than 70,000 enthusiastic volunteers for their dementia-focused studies and trials. Participants in Great Minds and the Clinical Studies Register are recruited from existing cohort studies, having undergone extensive characterisation, and from Join Dementia Research. Our interactive feasibility tools enable researchers to determine, quickly and easily, the number of volunteers that meet their study's requirements.

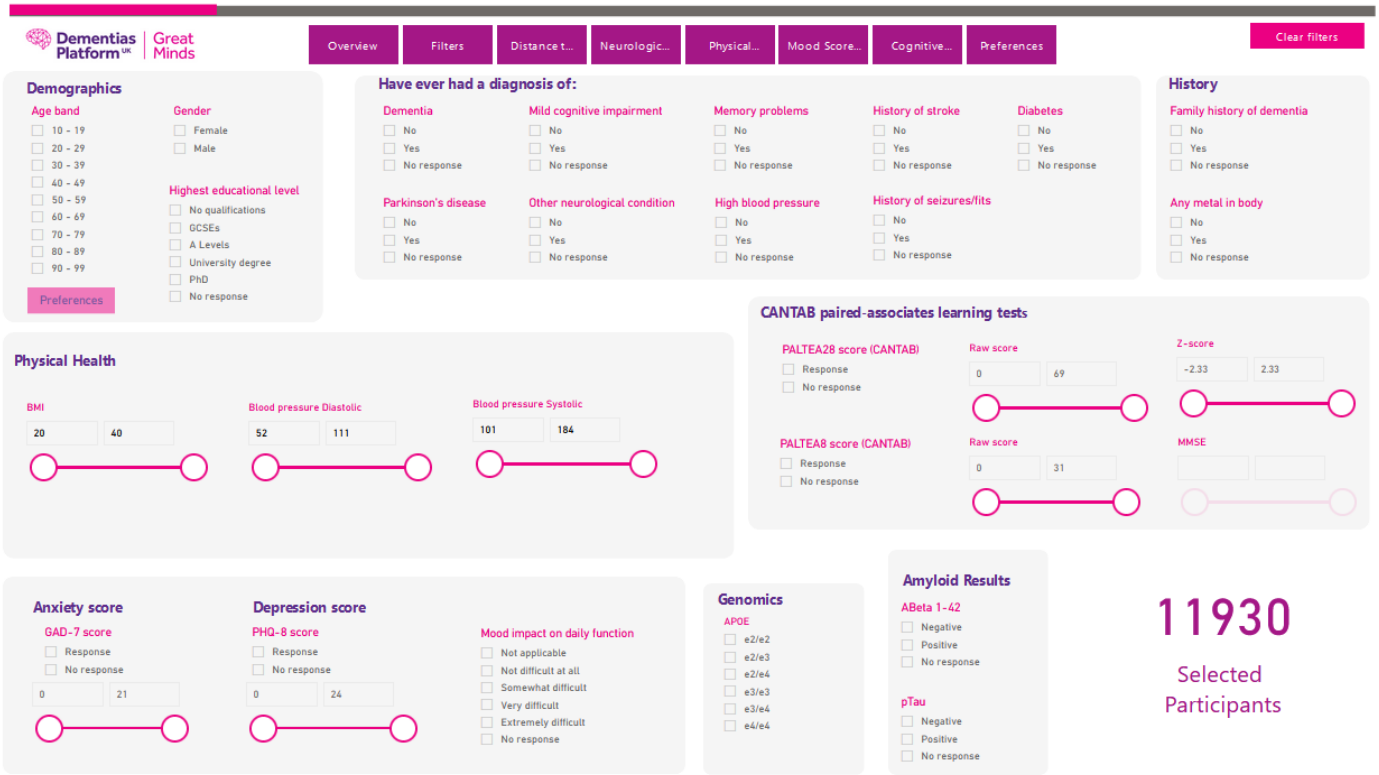
Dr Ivan Koychev

DPUK's Research Volunteer Registry Lead

"By working with volunteers from existing cohort studies, we can identify the people at risk of developing dementia symptoms and test the impact of new interventions in highly targeted trials."

Find the Right Participants for your Study

Access to the registers free. It is a rich resource that makes it significantly easier to recruit participants to a trial or study. Making it easier to do studies increases academic opportunities in dementia research and renews pharmaceutical industry interest in developing dementia therapies. Our interactive feasibility tools enable researchers to determine, quickly and easily, the number of volunteers that meet their study's requirements.



Artificial Intelligence Governance

Enabling the Development of AI within TREs

AI models can be prone to privacy attacks and data leakage which allow for inference or reconstruction of the training data used to develop them. There are also various characteristics in these models which can make them more vulnerable to these privacy attacks. Release of these models from a Trusted Research Environment poses a unique risk. DPUK Data Portal has developed a community of data providers, researchers and public representatives to review the risks involved in AI model release. Through this work, an AI model risk evaluation framework has been developed that has been deployed across DPUK to allow AI models to be securely developed and assessed. Further information about this work can be found at: AI_Risk_Report

AI Governance Processes

Our governance processes ensure that AI projects are done in a safe way that protects the privacy of the data they use for training. We have developed assessments and tools to help data owners make informed decisions when it comes to reviewing these projects.

1. AI Risk Training
2. AI Risk Assessment
3. Risk Evaluation Matrix
4. Privacy Attack Evaluations
5. Agreements & Licenses

AI Hosting

The DPUK Data Portal has enabled the ability to share, federate and deploy models so that we can facilitate AI research in a safe and secure way. This world-leading governance, coupled with our technological abilities, has allowed us to lead the way in AI research within TREs.

01 **Share**
Share the model via the TRE to enable open & reproducible science while keeping data secure.

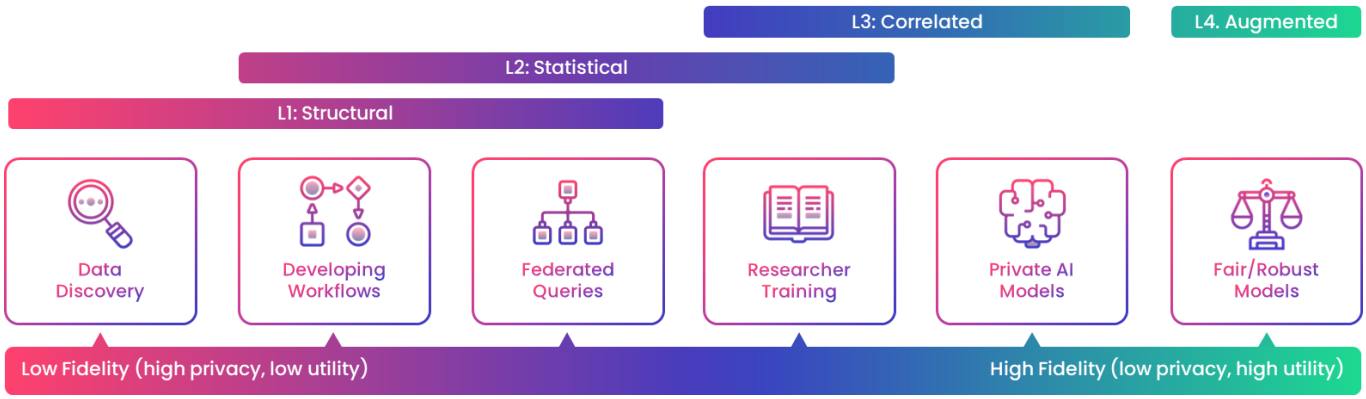
02 **Federate**
Federate between different data environments to enable fine-tuning or validation of AI models.

03 **Deploy**
Host the model via the TRE and allow it to be securely queried externally with restrictions.

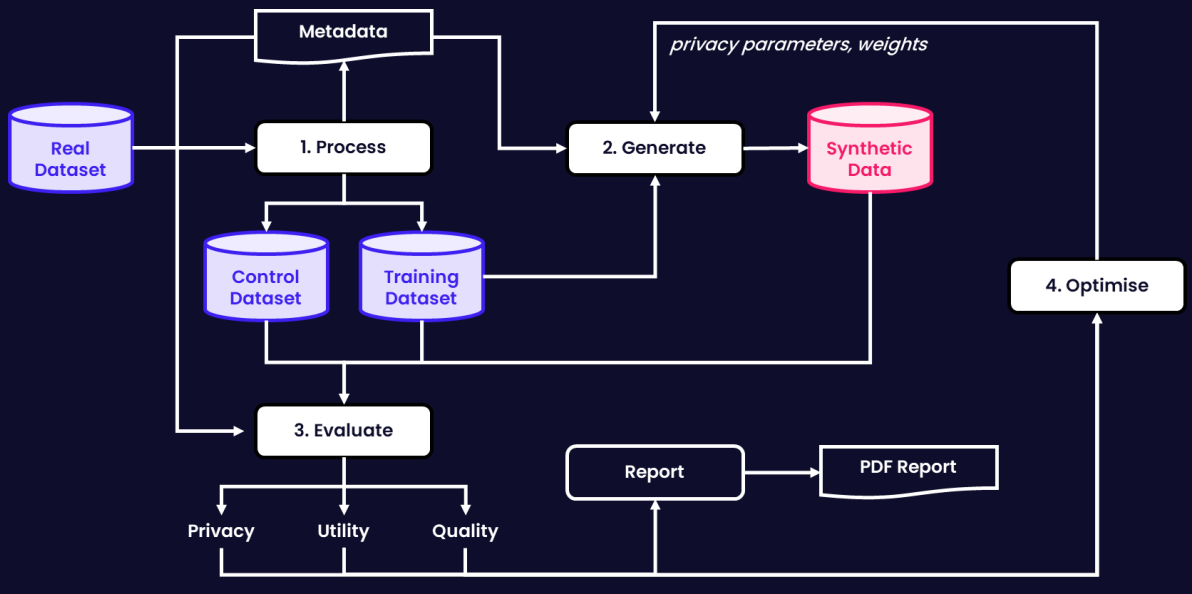
Synthetic Data

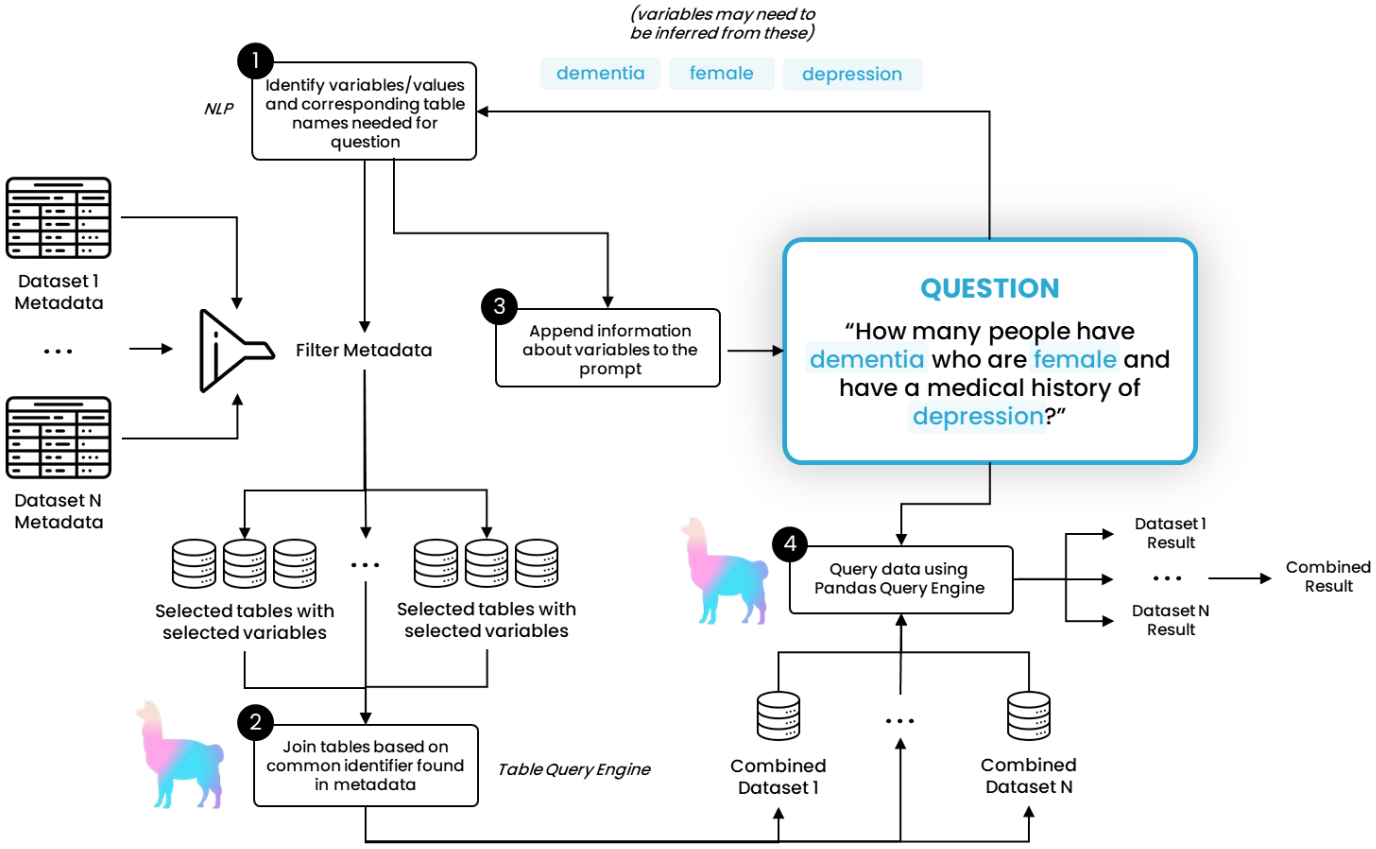
Providing Greater Access to Data

Synthetic data is artificially created data, typically created by generative models trained on real world datasets. Synthesiser models aim to learn the statistical properties, distributions and correlations of the real data, to be able to generate realistic artificial samples which are statistically similar to the data it was trained on.



Synthetic data has many different uses cases in sensitive health data research. For example, we can use synthetic data to enable researcher/student training in how to analyse health data, without the need for lengthy governance approval processes. It can also be used by researchers to develop their algorithms and workflows while waiting for data access. Additionally, we can use synthetic data in federated data scenarios where real data cannot be seen. DPUK has created an open-source Python package – SynthOpt – to be able to generate varying levels of synthetic data for these different purposes.





Large Language Models

Improving Data Discovery & Project Feasibility

DPUK is in the progress of implementing innovative generative AI solutions into the Data Portal to improve data discovery and project feasibility. With so many different cohort datasets within DPUK, each with thousands of variables, finding the right datasets can often be challenging for researchers. We aim to streamline this process by enabling researchers to query our data with natural language to identify which datasets contain the data that is needed for a given project. It will also allow researchers to dig deeper into the data to find out whether there are enough participants with a certain disease, or if the data contains other modalities such as neuroimaging and genomics.

This new innovative approach allows researchers to go beyond the typical data discovery and metadata methods to allow a deeper understanding of the datasets before applying for access.

Dementias Platform UK

Powered by SeRP

ISO certified and UK Statistics Authority Accredited

portal.dementiasplatform.uk