

BHF Data Science Centre Cardiovascular and Diabetes Cohorts Trusted Research Environment (TRE) call: response to queries received (21/12/2022)

We have had some queries from organisations potentially interested in applying to this call.

The initial aim is to first ensure that linkage to nationally collated health data can be achieved for the relevant cohorts, with support for analysis across complex multimodal data types being a later aim. We would therefore prioritise this at the start of the programme and are having ongoing discussions with NHS data custodians across the devolved nations to deliver this.

Although federated solutions are of potential interest, because of the logistic challenges of implementing these in partnership with multiple research organisations, we had envisaged a TRE into which (initially mainly structured) data from cohorts and national health data will be ingested and linked. Such an environment would allow us to ingest other (complex and unstructured) data types including from omics analyses, imaging and wearables. It is envisaged that cohorts will then be able to carry out their analyses across all data modalities for scientific insights.

Solutions for different requirements could be provided/proposed by a single organisation. However, we will consider assembling solutions proposed by organisations with particular expertise or capability in different areas/ data types provided it is possible to facilitate appropriate partnership agreements between those organisations, either as part of this particular funding call or – if appropriate - with future funding.

Solutions can include 'cohort discovery' tools, although our key focus is on ingestion and linkage in the first instance, with these tools coming in later where appropriate.

The funding available for the call is £580,000 and bids up to and including this value are expected. Provided the organisation can capitalise the asset, the available funds can be used across 3 years and are expected to cover the costs of all elements put forward by the proposing organisation, including storage and compute, where appropriate. There may be areas where the funds available could be supplemented if required and justified.

The following information is provided as a guide to calculating storage costs: as an example, if we assume for our cohorts that all elements of data are available, then this would be as follows:

- Structured bespoke clinical and lab result data from cohort
- NHS linked data- usually all structured data; includes e.g. Hospital episode statistics, GP data, medicines data, emergency admissions
- Registry data e.g. stroke and myocardial infarction national audits- all structured
- Proteomics/ metabolomics- usually flat files
- Genomics- sequence or genotyping, including all processing
- Imaging (e.g. via XNat connection, can be unstructured in the first instance)
- Wearable data

For these we would estimate a storage requirement of **1Gb per participant** to ensure processing requirements of some of the elements of the data (e.g. genomics and imaging in particular if raw data are ingested). Data are currently stored by individual cohorts within parent research organisation data safe havens.

Cloud storage is acceptable however we require that this is UK-based (due to national data custodian requirements) and these costs can be included in the bid.