UK Biobank Managed Informatics Platform for Research Data Access Response to Prior Information Notice

Introduction

UK Biobank is a large-scale prospective study of ~500,000 UK volunteers that aims to improve the prevention, diagnosis and treatment of a wide range of serious and life-threatening illnesses – including cancer, heart disease, stroke, diabetes, arthritis, osteoporosis, eye disorders, depression and dementia. UK Biobank is a de-facto open-access resource: there is a clear access test which is that the researcher has to be a bona fide researcher and the research has to be health-related and in the public interest, and all researchers and applications are subject to the same access criteria.

Over the next five years, it is anticipated that the UK Biobank resource will comprise over 15 Petabytes of data based on current scientific programmes. In the medium term, the UK Biobank resource will likely be further enhanced (although there are no firm plans as yet in place) by various large-scale –omics assays (such as proteomics and/or metabolomics).

UK Biobank therefore requires a new approach to data storage and access, which allows researchers to bring their analyses to the data through implementation of an access and analysis informatics platform.

UK Biobank published a Prior Information Notice relating to a pre-procurement market awareness phase, outlining its considerations for a managed informatics platform accompanied by a set of questions against which it welcomed feedback from potential suppliers. A link to the Prior Information Notice and the supporting overview and questions document can be found on the UK Biobank website:

https://www.ukbiobank.ac.uk/managed-informatics-platform/

This document summarises responses received in response to the PIN notice.

Response to Prior Information Notice

A series of videoconferences were held during November 2019 between UK Biobank and suppliers who had responded to the UK Biobank PIN notice.

Suppliers responding to the PIN notice included those with an established platform that supported genomic and other analyses, those who proposed to develop a solution based upon more general database and query technologies, and those who could provide cloud infrastructure or specific technology components.

The responses submitted together with subsequent videoconference discussions confirmed that the capabilities UK Biobank is seeking to procure do exist in the market place, although the maturity of specific capabilities varies from supplier to supplier.

There was general consensus on a number of key points:

- *A public cloud infrastructure would provide the most appropriate base for the platform;*
- *Clear definition of the requirements was needed, especially for a managed service;*
- *The timescales whilst challenging were achievable; and*
- *A proof of concept demonstration using synthetic data would be of benefit.*

Functional Capabilities

Currently, researchers access UK Biobank data by downloading them and working with them in their own environment. The proposed platform will allow users to work with data in situ, and provide features such as a cohort browser and analytic tools and pipelines available as part of the platform.

Such a platform approach will also help democratise data access, lowering the barrier of entry to allow access by researchers from low economic development countries or from institutions which do not have the budget to spend on storage or compute infrastructure at scale.

A cohort browser is needed which will allow researchers to explore the cohort, and select a sub-cohort based on phenotypic, genotypic, or other characteristics for further analysis.

Access to UK Biobank data held by the platform will be controlled by the same mechanisms as it is today. UK Biobank will provide API access to its AMS system to allow researchers to be authenticated and authorised for the appropriate level of data access. API provision of project specific pseudonymised identifiers and a mapping file for data held by the platform will also be provided.

A number of suppliers could offer cohort browser capability, either as existing functionality or through front end development. Support for access via Jupyter notebooks, R Studio, and other user interfaces are readily available.

Additional tools and pipelines could typically be supported via e.g. Docker containers and a workflow description language such as CWL or WDL.

Scope of Data

The initial focus of the platform will be on genotypic and phenotypic data, but will also need to support additional data types already available such as imaging and activity monitoring data as well as new data types that may emerge from future projects.

The platform will need to ingest a full copy of the UK Biobank dataset, including both structured and 'blob' data. These data will not make use of proprietary formats.

Linked healthcare records (both primary and secondary) will be made available within the platform; the linkage will be undertaken by UK Biobank separately prior to ingestion of these data by the platform.

The data that UK Biobank makes available is de-identified, and UK Biobank therefore has no preference currently as to data residency; however, as data protection regimes may change over time it needs to retain the ability to specify this in future.

Suppliers are able to handle both structured and blob data, though with differing capabilities depending on the focus of their prior implementations.

Infrastructure

UK Biobank do not have a preference for a cloud provider, and will want bidders to suggest the best option that they can offer, rather than offer several alternatives with different cloud providers. The

platform needs to be transferable to support potential future extension to multiple platforms or multiple cloud providers.

Most suppliers viewed public cloud as being the most appropriate underlying infrastructure. Most suppliers were cloud agnostic, and able to implement their platform on a variety of public cloud instances.

Managed Service

The platform must be delivered as a managed service, presenting a single face to researchers even if several parties are supporting it in a consortium. It must be able to meet the needs of different researchers, from a range of disciplines and with differing levels of technical expertise.

All suppliers could offer a managed service, with the main emphasis being on platform technical support for end users; some also covered operational support including e.g. data ingestion. User support was typically provided by online help and user forums, as well as via a more traditional help desk. A number of suppliers noted the ticketing systems they used (Zendesk, Jira), and the potential for extended use by UK Biobank staff.

Several suppliers highlighted the importance of defining the 'service wrap' requirements in sufficient detail in any tender. These would include how support should be shared across supplier and UK Biobank teams (e.g. L1/L2 help desk), the areas in which such support would be required (e.g. platform capabilities, scientific advice, data queries), and call handling flow.

Billing and Charging

Costs related to compute and additional data ingress/storage/egress need to be clearly identified, and be charged directly to researchers by the platform provider (not via UK Biobank).

Whilst all UK Biobank data will need to be in a high performance storage environment initially to allow 24/7 access to researcher in over 70 countries, over 2-4 years the residual value in the raw data will reduce and such data could potentially be moved to a more cost-effective storage environment over time.

Most suppliers could attribute resource consumption costs to individual projects, typically leveraging underlying cloud infrastructure capabilities. Many were able to take on billing responsibility rather than simply pass back a summary of charges incurred to UK Biobank for recovery from researchers.

Commercial Model

UK Biobank would be interested in commercial models that enable it to reduce the cost to itself of core platform provision, provided these did not significantly increase the costs incurred by researchers wishing to use such a platform service and maintains non-preferential access.

Suppliers suggested a number of commercial models that might be considered, including:

- *Charging more granular access fees for the various types of data in the resource;*
- *Differential charging for academic and industry researchers;*
- *Leveraging the public dataset hosting offered by some cloud infrastructure providers; and*
- *Charging researchers market rates for resource consumption, rather than the lower rates achieved by the scale of overall platform consumption.*

Timeline

The timeline for platform procurement and implementation is expected to be:

- Formal tender go-live in December 2019;
- Tender responses in January 2020;
- Demonstration in February 2020;
- Contract in place in March 2020;
- Beta platform in place in April 2020; and
- Production pilot service in May 2020.

UK Biobank is not looking for component parts, but rather an end-to-end managed solution with all infrastructure included. It is envisaged that the solution will need to be delivered largely with existing capabilities to meet the required timeline constraints, rather than involving extensive bespoke development.

Almost all suppliers considered the timescales challenging but achievable, based upon a significant degree of existing out of the box functionality. A few who needed to undertake more extensive bespoke development anticipated that a Q2 2020 beta and Q3 2020 go-live as more appropriate.

Key risks identified included unclear requirements, the need for custom user interface, and researcher pseudonymised identifier mapping. Key dependencies on UK Biobank related to information about, and the availability of, APIs to support e.g. authentication and authorisation.

The contract awarded from the tender round will be for 2-3 years initially, to allow for the possibility that UK Biobank will eventually need multiple platforms and/or multiple cloud providers, for which further tenders may be run.

Most supplier saw a minimum term of 2-3 years as being appropriate with the option to extend.

Enabling technologies

A number of suppliers referenced specific technical capabilities provided by their public cloud provider of choice (e.g. FGPA, GPU support). Some suppliers offered specific technologies for e.g. genomic data compression, encryption, and controlled access.

Proof of concept demonstration

During the call, UK Biobank outlined its desire to include a demonstration/proof of concept exercise as part of the tender evaluation process. This would involve providing a synthetic dataset at the time of tender release, and outlining a number of scenarios (e.g. browsing the cohort, use of project specific participant pseudonymised identifers, running various types of analysis) that suppliers would be asked to demonstrate. The synthetic dataset would represent the richness of the UK Biobank data, but not the volume.

All suppliers welcomed such an exercise, seeing it as both an opportunity to demonstrate their capabilities and as a way de-risking an accelerated implementation following contract award.

Clarification questions raised

Q: How many individual researchers will there be at the initial stage of the platform?

A: The initial implementation will be for a small set of defined users; incremental functionality and data types, and additional research groups will follow.

Q: What areas will be scored in the tender, and what will weighting be?

A: Tenders will be scored across a number of areas, possibly including:

- Informatics functionality (current and planned);

- Support functionality e.g. billing;
- Audit;
- Data ingestion;
- Service quality;
- Demonstrable ability to implement; and
- Cost.

Scoring and weighting will be set out clearly in the tender documents.

Q: Will the platform need to allow researchers to transfer between projects within the platform, with different user names and passwords?

A: Since different research projects have different participant pseudonymised identifiers, researchers will need to re-authorise/re-authenticate to work on different projects.

Q: How will UK Biobank characterise users, and define their user experience?

A: This will be considered and included in the specification.

Q: Do UK Biobank have a specific country the data must be stored in?

A: Whilst there is no specific geographic requirement, there is a preference for an EU country.

CB/AH/CC 20/11/2019