

The Barts Health Data Platform

October 2024

Authors: Ruzena Uddin, Benjamin Eaton, Steven Newhouse, Evan Hann, Tony Wildish, and Idowu Bioku

Barts Health NHS Trust, Ashfield Street, Whitechapel, London, E1 2BL, UK

bartshealth.researchdata@nhs.net

1. Abstract

The Barts Health NHS Trust (BHNT) has established the Barts Health Data Platform (BHDP) as part of a broader Precision Medicine Programme (PMP) to support research into, and operational deployment of, advanced data analytics for the 2.5M patients and public primarily living in East London. The BHDP consist of three new components: the Barts Health Data Access Committee (BHDAC) (which brings together relevant professionals and public to review and make data access decisions), the Analysis Data Core (ADC) (which transforms raw patient data from various internal sources into a coherent data set for onward analysis) and a Secure Data Environment (SDE) (where authorised projects and users analyse sensitive data). This paper will provide a high-level overview of the design and implementation of these components, and how they provide the governance and technical measures needed to successfully deliver secure access to sensitive health records. This work has been supported by Barts Charity¹ through two awards that will enable the design and operation of the BHDP between 2021 and 2027.

2. Background

2.1 Sensitive Health Care Data Analysis

Over the last five years, sensitive data within the UK is increasingly being held within Trusted Research Environments (TREs) where approved researchers can be provided with access to the specific data they have been authorised to access. This is a move away from the historical approach in some communities where sensitive data was given to authorised individual researchers under contract to analyse the data in their home institutions and environment. Using a library analogy this change represents a move away from the previous 'borrowing' model where data is taken away, to a 'reading room' model where users are

¹ Grant awards G-001725 and G-002196.

required to do their work in a self-contained environment and are limited in what they can take in or out.

This change allows for greater flexibility, allowing linked, de-identified data sets to be securely provided to answer vital research-related questions. The TRE model itself is now being evolved to a SDE model reflecting a greater reliance of technical and organisational measures to ensure security and integrity rather than just relying on trust. The current generation of SDEs and TREs provide a data and research analysis platform, that can uphold the highest standards of privacy and security of health data when used for research and analysis which patients and the public are expecting.

The need for SDEs or TREs within the NHS is driven primarily by the increasing digitalisation of the NHS and the growing availability of patient data through the Electronic Patient Records (EPRs) and related data sources (i.e., images) for research and clinical purposes.

2.2 The Need at Barts

BHNT encompasses 5 hospitals in East London covering over 2.5M current patients with an integrated EPR which has been operating for over a decade containing a total of 8M patient records. While the digitalisation and subsequent integration of our processes continues, our main data warehouse has reached 5PB² and continues to grow and already provides an integrated raw source of data for onward data analysis. However, supporting such analysis for research purposes sits outside the remit of operational NHS staff meaning that time and effort to support the broader use of health data inside and outside the hospital was always challenging to prioritise and justify.

Barts Charity works closely with BHNT to improve the health care for the people of East London. Strategically, it was recognised by both parties that improving access to data for research and clinical purposes would be a key enabled to better understanding the health challenges in East London – both at a population level and individually through personalised or precision medicine. In early 2020 Barts Charity funded the PMP with Barts Life Science - a partnership between BHNT and Queen Mary University of London (QMUL). The PMP had two key strands: the design and implementation of what has become the BHDP (led by BHNT) and the establishment of three research groups in QMUL that would use the data platform as one of their data sources.

Due to the pandemic, what would become a 6-year programme did not start until 2021 with the BHNT team initially undertaking a scoping study to better understand the key

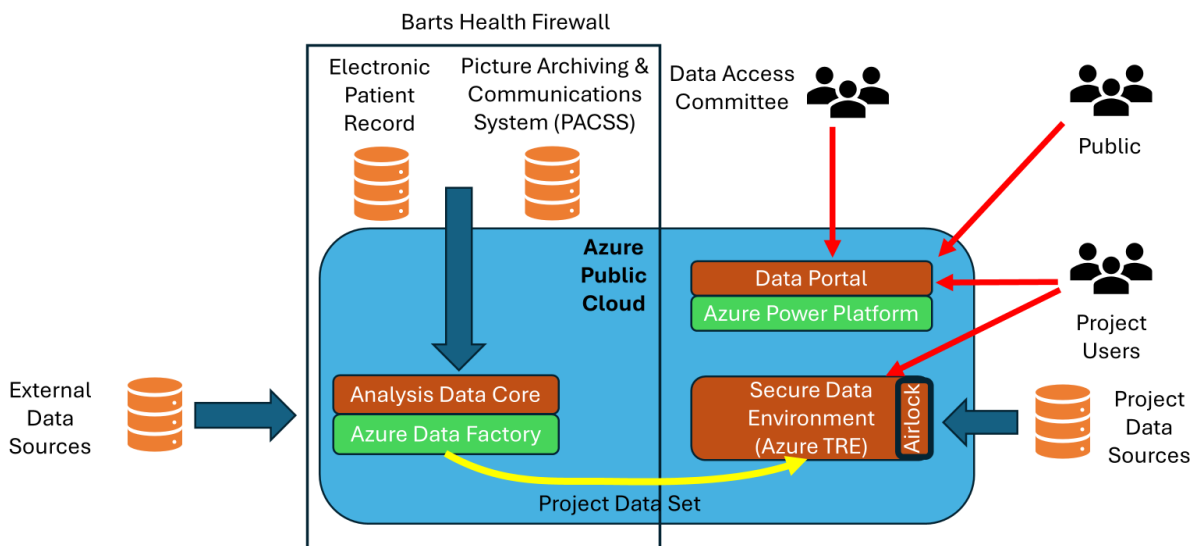
² One Petabyte is the equivalent of 20 million tall filing cabinets or 500 billion pages of standard printed text or around 200,000 single layer DVDs.

components and workflows around the BHDP. As a result of this scoping study, a further implementation project that would design and build what would become the BHDP was developed and successfully funded by Barts Charity for 4 years until 2027.

3. Barts Health Data Platform

The BHDP has three key components that have been established to deliver:

- Support for users to make a Data Access Request through a public Data Portal and robust governance to ensure compliance of process and legislation through the BHDAC.
- Integration and transformation of patient data ready for presentation to individual analysis projects through the ADC compressed to 4TB covering 10,000 data points.
- A SDE where users can analyse the data, they have been given access to in a restricted workspace.



The individual design requirements and operation of these three components are described in the following sections. Given the existing investment that BHNT had made in migrating on premise infrastructure to Azure, there was a strong expectation that the BHDP would be based upon Azure to benefit from the institutional investment that had already been made.

Additional expertise was sourced into the team to build out the initial stages of the BHDP through a public tender launched using the GCloud –13 framework which completed in January 2024. From the offered solutions, PA Consulting were selected to assist the team in

this startup phase. The work involved building the Data Portal within the MS Power Platform environment and providing additional technical expertise around the SDE deployment using the Azure TRE as a starting point.

3.1 Barts Health Data Access Committee

A recurring theme in talking to researchers, clinical data users and internal data providers during the scoping phase of the project was the difficulty in finding out the correct processes that were needed to obtain authorisation, the difficulty in finding the right person or team to give you the data (or even to find out what data was available), and finally getting clarity that approval had been given. There was also an organisational desire to gain visibility on all the large-scale uses of patient data taking place within the hospital – hence a need to bring together both the research and clinical data requests – and to integrate with the existing internal approval workflows.

Our solution is focused on a Data Portal that provides a public ‘front door’ for triggering a request to access patient data for analysis at the hospital for research or clinical use. The portal provides information around the data and related meta-data held by the hospital, the projects already being supported by the hospital’s data, a web form and supporting workflows to handle the approval process ensuring all stakeholders are informed of the request status, operation of the DAC, and the engagement that is taking place with the patients and public in East London.

The BHDAC was established in January 2022 to provide a single decision point through which all requests to access patient data at scale for research or clinical purposes. Comprised of members of the Precision Medicine team, professionals from Information Governance and Research Governance, representatives from the local Research and Clinical communities, the BHNT commercial team and public contributors representing the public and patients in East London. Generally, meeting monthly the DAC undertakes a review of all proposals primarily focusing on the relevant ethical, legal and technical aspects of the request to ensure our compliance with all regulatory processes. The pre- and post-DAC processes allows us to enforce the ‘Five Safes’, additional criteria introduced by NHS England and the relevant guidance in the SATRE specification and other processes.

A Patient and Public Involvement and Engagement (PPIE) strategy has been developed and approved by the programme board which includes input from public contributors on the board and the Barts Health Engagement team. In line with our PPIE strategy, we have appointed 3 public contributors to the Programme Board and 2 public contributors to the DAC to provide a broader perspective and input on our decision making. In July 2024, 6 members of the public were recruited to participate in a focus group to discuss the visual

identity of the BHDP. Members of the public were able to contribute their perspective and help shape an important aspect of our program from the public's view. The visual identity will be featured on our website, documents, and merchandise.

3.2 Analysis Data Core

The ADC provides a secure cloud-based environment within Azure, logically behind the Barts Health firewall, where patient data is brought together from various sources across the hospital using the Azure Data Factory (ADF), a fully managed scalable data integration environment provided within Azure, with regularly scheduled daily pipeline runs into a secure Databricks³ instance.

The ingestion of patient data from the hospital triggers various Databricks pipelines which work to combine the data sets with each other and with external reference data, before performing various cleaning and checking functions to produce various data sets for onward use⁴. These pipelines leverage delta live tables and Apache Spark in the Azure cloud to request compute when required, run transformations in parallel and do incremental loads to maintain up-to-date data products.

Project specific fields are extracted from these raw or transformed data products based on the needs of the individual project. If a project requires data which is not presently in one of the current data products this data is identified from within the hospital sources and brought into the pipeline. Fields are categorised according to their Information Governance risk to ensure that the data extracts do not contain any data beyond the approvals of a given project.

For anything other than anonymised data to be provided to the project, we would require the project to provide additional ethical consent or for the field to undergo an additional deidentification process. If the data extract is going to be combined with other data sets, then we will work with other data controllers to generate a data extract with pseudo anonymised identifiers and perform data linkage where required.

The ADF infrastructure is used to push data securely into the SDE to allow a project to access it. These pipelines can be manually triggered or configured to provide regular automatic refreshes of the data as desired by the project.

3.3 Secure Data Environment

The SDE provided within the BHDP provides a secure workspace for each project to access patient data which they have been authorised access to. Following an analysis of the

³ <https://www.databricks.com/>

⁴ <https://github.com/Barts-Life-Science/Research-Data-Extract>

commercial ‘As A Service’ and open-source offerings we selected the Azure TRE⁵ as the most mature available open-source offering (Spring 2023) that would provide us with the most cost-effective long-term flexibility that could be operated by a small internal team. We felt this encapsulated the right level of flexibility between procuring a turn-key system where any adaption or extension of the base offering would come at extra cost and negotiation, versus the development risk of building our own solution from the ground up with a small budget-limited team. All the development work being undertaken locally to make the Azure TRE deployable for our production purposes is publicly available⁶ and will be contributed where relevant to the upstream repositories.

The initial trial deployments (alpha) in Summer 2024 were used to gather user feedback from over 15 researchers on the offered features without access to patient data. This environment allowed us to prioritise the next phase of developments and allow independent penetration testing to take place. Missing functionality included a reliable process to build a suite of VM images or to deploy and redeploy the Azure TRE in stages. Standard deployments (on the early versions we deployed) included security issues such as a default network configuration that allows data to be exported from the SDE using DNS calls, a default virtual desktop configuration that allowed audio to be exported from the project workspace and unlimited text to be cut and pasted into and out of the SDE. Most of these issues are relatively easy to resolve, but underline that the out of the box Azure TRE distribution can only be considered for internal use.

With the critical issues in the alpha addressed we moved to an invited beta deployment in September 2024 that provided access to the relevant patient data for the authorised projects. The beta phase lasted until November 2024 and provided time to validate the patient data workflows and enhanced user environments, allowing us and our users to gain confidence in the SDE. We plan to move to an open Minimal Viable Product (MVP) release before the end of 2024.

4. Current Status

As of Autumn 2024, the SDE has been in development for 9 months and has moved from initial trial deployments, through alpha and beta releases, before being launched internally as an MVP. The beta phase will include further penetration testing to validate the changes made as a result of our earlier analysis and to build confidence in the MVP before the public launch in 2025.

⁵ <https://github.com/microsoft/AzureTRE>

⁶ <https://github.com/Barts-Life-Science/AzureTRE>

While the investment from Barts Charity has helped the launch of the BHDP, there is a need for an income stream to be established to ensure sustainability by covering operating and future development costs. A three-tier charging model is being envisaged covering internal usage with charges set essentially at costs with an overhead (Tier 0), a bundle for small scale external projects (Tier 1) and a bundle for larger scale external projects (Tier 2). Additional revenue clauses could be added to project agreements depending on the exploitability of the data – either through up front charges or downstream revenue clauses. In all cases data will remain the property of BHNT and control will be maintained on how the data is exploited through these agreements.

5. Future Plans

We expect the MVP phase to run for 3-6 months in early 2025 while we complete further user driven developments and complete the integration of the SDE into the BHNT's security certification, with the launch of the complete platform in mid 2025. This would allow us to launch fully integrated into the NHS IT security landscape, a rich diverse set of patient data encompassing patient health care data and images from our PACS.

6. Acknowledgements

We would like to thank Barts Charity funding through awards G-001725 and G-002196 for supporting this work.

Our early adopting user community for their support and feedback through the QMUL research teams supported through Barts Charity within the Precision Medicine Programme and the Barts Life Science Data Science Team.

The support and engagement of those we have worked closely with the team over the years: Sarah Jensen (Chief Information Officer, BHNT), Chales Gutteridge (Chief Clinical Information Officer, BHNT), and Jennifer Dando (Deputy Information Governance Lead, BHNT).