

Article

Dynamic Biobanking for Advancing Breast Cancer Research

Maryam Abdollahyan¹, Emanuela Gadaleta¹ , Millahat Asif¹, Jorge Oscanoa¹, Rachel Barrow-McGee², Samantha Jones², Louise J Jones² and Claude Chelala^{1,*}

¹ Centre for Cancer Biomarkers and Biotherapeutics, Barts Cancer Institute, Queen Mary University of London, London EC1M 6BQ, UK

² Centre for Tumour Biology, Barts Cancer Institute, Queen Mary University of London, London EC1M 6BQ, UK

* Correspondence: c.chelala@qmul.ac.uk

Abstract: Longitudinal patient biospecimens and data advance breast cancer research through enabling precision medicine approaches for identifying risk, early diagnosis, improved disease management and targeted therapy. Cancer biobanks must evolve to provide not only access to high-quality annotated biospecimens and rich associated data, but also the tools required to harness these data. We present the Breast Cancer Now Tissue Bank centre at the Barts Cancer Institute as an exemplar of a dynamic biobanking ecosystem that hosts and links longitudinal biospecimens and multimodal data including electronic health records, genomic and imaging data, offered alongside integrated data sharing and analytics tools. We demonstrate how such an ecosystem can inform precision medicine efforts in breast cancer research.

Keywords: breast cancer; translational research; precision medicine; electronic health records; genomics; health informatics; bioinformatics



Citation: Abdollahyan, M.; Gadaleta, E.; Asif, M.; Oscanoa, J.; Barrow-McGee, R.; Jones, S.; Jones, L.J.; Chelala, C. Dynamic Biobanking for Advancing Breast Cancer Research. *J. Pers. Med.* **2023**, *13*, 360. <https://doi.org/10.3390/jpm13020360>

Academic Editors: Carlos Martínez-Pérez and Arran K. Turnbull

Received: 23 January 2023

Revised: 15 February 2023

Accepted: 16 February 2023

Published: 18 February 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Approximately one in eight women are diagnosed with breast cancer in their lifetime [1]. Despite recent advances in early detection and treatment of BC, and extensive histopathological and molecular characterisation of primary tumours, around 30% of patients go on to experience recurrence [2] or metastasis which is considered incurable. These figures are expected to rise in the upcoming years due to the disruptions in screening programmes and treatments related to the COVID-19 pandemic, with nearly 1 million women in the UK having missed their mammogram appointments [3]. While effects of the COVID-19 pandemic amplified existing challenges and health inequalities, increasing demand for innovative solutions, the risk factors for breast cancer remain not well-understood. This is particularly evident in multi-ethnic communities who are underrepresented in or absent from large studies. Cancer biobanks are important biomedical research infrastructures for preclinical, translational and clinical studies that can help address these challenges.

As we move forward in the era of precision medicine, cancer biobanks are required to adopt new strategies for patient recruitment and collection, processing, storage and distribution of biospecimens and data to better contribute to this research area and the wider healthcare landscape. Linkage to other data sources such as electronic health records (EHRs) is vital to complement a biobank's clinical data. Healthcare datasets, as managed in EHR systems, are typically stored in highly variable formats. There is homogeneity in terms of structure and adoption of standard biomedical ontologies (e.g., ICD [4], OPCS [5] and SNOMED-CT [6]); however, a notable portion of EHRs (e.g., discharge summaries) is in a free-text format. While these unstructured data are often used to report the results of examinations, tests or procedures, and serve as a means of communication between healthcare providers, they contain valuable information that is not possible to obtain from their structured counterparts. Such information can be extracted using artificial intelligence

algorithms and stored in a structured format for subsequent processing. Similarly, when combined with a biobank's clinical data, sequencing data allow researchers to discover new variants that contribute to disease onset and progression, identify drug targets for accelerating drug discovery and development, and stratify patients into groups such as those more likely to respond to treatment or experience side-effects. Moreover, provision of tools for analysing and sharing these data is crucial to maximise the use of the biospecimen and data hosted by cancer biobanks.

Breast Cancer Now Tissue Bank (BCNTB) [7] is the UK's first national breast cancer biobank, providing access to the largest collection of longitudinal biospecimens and data from breast cancer patients and individuals without the disease at various stages of the care pathway. The Bank has eight (four legacy and four active) centres in England and Scotland, with Barts Cancer Institute (BCI) acting as the coordinating centre. In this paper, we explore the biospecimens, data and tools made available by the BCNTB centre at BCI (BCNTB-BCI). We present an approach to building and maintaining a dynamic biobanking ecosystem, that features: a centralised database with built-in quality assurance mechanisms for storing multimodal data; linkage to data from external sources, including EHRs from primary and secondary care plus genomic data from national and international sequencing initiatives, for enriching the Bank's data; and an integrated biospecimen and data request system and a bioinformatics portal for sharing and analysing these data. While individual elements of such an infrastructure have been suggested in the past (e.g., see [8–10]), no cancer biobank has implemented all these elements, specifically for breast cancer research. Using BCNTB-BCI as an exemplar, we demonstrate how dynamic cancer biobanks can fill the gaps in breast cancer patient data, support researchers to advance breast cancer research and assist clinicians in early diagnosis and effective treatment of breast cancer for optimal outcome.

2. Materials and Methods

Donor Selection and Recruitment. Necessary ethics and governance approvals (NHS Research Ethics Committee reference number 21/EE/0072) are in place for the routine collection of longitudinal biospecimens and data from the donors attending breast cancer clinics across Barts Health NHS Trust (BH). They include patients undergoing biopsy or surgery for benign, atypical or malignant breast lesions as well as individuals undergoing surgery for cosmetic or risk reduction reasons. We check the National Data Opt-out list to exclude donors who have opted out from the use of their data for research. The Bank's consent form and patient information sheet are discussed in detail with eligible donors, giving them the opportunity to ask questions. Donors are informed that they can withdraw their consent at any time. Following consent, each donor is assigned a unique biobank ID which links their biospecimens to the associated data.

Biospecimens and Multimodal Data. BCNTB-BCI hosts a repository of biospecimens and multimodal associated data. These biospecimens include liquid samples (e.g., whole blood, plasma, serum and buffy coat), fresh and frozen tissue samples (e.g., tissue microarrays (TMAs), formalin-fixed paraffin-embedded (FFPE) blocks and haematoxylin and eosin (H&E) stains) and primary cells isolated from breast tissue (e.g., epithelial and myoepithelial cells and fibroblasts). Tissue samples (tumour, cancer-surrounding, benign or healthy) are generally surplus to diagnostic requirements, while liquid samples and derivatives are taken specifically for the Bank. All biospecimens are collected and prepared according to BCNTB standard operating procedures (SOPs). The associated data are manually curated by the Bank through a variety of sources (e.g., patient questionnaires and the East London Patient Record (eLPR) system [11]). These data include donor demographics and survival status, personal and family medical history, lifestyle, presentation and diagnosis, procedures (biopsy, surgery and imaging), treatments (chemotherapy and radiotherapy), follow-up reports and images (scanned H&E slides) manually extracted from clinical letters, pathology and imaging reports and vital sign readings.

Data from BCNTB-BCI and all other BCNTB centres are stored in a centralised Research Electronic Data Capture (REDCap) [12,13] database, a secure web-based data capture application for building and managing online surveys and databases to support research studies, at the Bank's coordinating centre. Since BCNTB is a multisite national biobank, we defined a core data dictionary that allows shared data processing and reporting across sites. Data from all the legacy BCNTB centres were restructured to adhere to this data dictionary. Several REDCap features are utilised to enable and improve multisite access, data entry and validation (via branching logic, calculated fields and custom scripts run through the REDCap API), audits (e.g., via custom reports) and regulatory compliance. BCNTB-BCI data are manually audited periodically, with each data point being verified by at least two members of the Bank. Furthermore, BCNTB-BCI data are mapped to concepts from the SNOMED-CT biomedical ontology to facilitate the exchange and comparison of these data with information from other sources (e.g., EHRs and genomic data). Any data distributed for research are pseudonymised.

Linkage to EHRs and Genomic Data. To supplement and validate the Bank's manually curated internal data, we considered linkage to external data from primary and secondary care as well as national and international sequencing initiatives. BCNTB-BCI data are linked to EHRs from BH via the NHS or hospital number. BH is the largest NHS trust in London with five hospitals (St Bartholomew's, The Royal London, Mile End, Newham and Whipps Cross) serving 2.5 million people across East London [14]. These EHRs include patient demographics and survival status, medical history, physical and social assessments, diagnosis, procedures, imaging reports, pathology results and histopathology reports, electronic prescriptions, maternity services data, clinician notes and cancer registries data (e.g., the Somerset Cancer Register [15]) that are consolidated from several datasets hosted or linked by BH [16]. To transform EHRs from BH into clinically relevant features for construction of cohorts suitable for breast cancer research, we performed a series of data cleaning and wrangling processes. During these processes, EHRs were assessed with regards to their syntax (e.g., format), schema (e.g., data model and structure) and semantics (e.g., clinical interpretation); duplicates, missing values and discrepancies were flagged to be manually reviewed and compared to information from other sources. In addition, we used natural language processing (NLP) techniques to extract a variety of information from the unstructured part of these EHRs, and stored it in a structured format (e.g., mapped to the SNOMED-CT concepts). This information includes mentions of cancer-related and non-cancer-related symptoms, conditions, drugs and outcomes. For instance, we applied NLP to over 45,000 free-text reports of various imaging procedures (e.g., mammogram, computed tomography (CT), magnetic resonance imaging (MRI) and ultrasound (US)) to extract information that can be used to predict breast cancer recurrence (Abdollahyan et al., manuscript in preparation).

To date, tumour and normal samples from many BCNTB donors have been sequenced and the results made available to researchers through repositories such as the European Genome-phenome Archive (EGA) [17]. Moreover, BCNTB-BCI data are linked to sequencing data from the 100,000 Genomes Project (100K GP) [18] via a unique ID for over 320 BCNTB donors who are dually consented by BCNTB-BCI and 100K GP. These data, where available, include primary clinical data for the 100K GP participants and secondary clinical data received from NHS Digital and the National Cancer Registration and Analysis Service [19].

Biospecimen and Data Request System. The Bank's expression of interest (EOI) and application forms are implemented in REDCap. The forms are accompanied by Sample Finder, a tool implemented using an open-source Python web development framework [20] that allows researchers to search and filter biospecimens by donor and tumour characteristics. Data shown in Sample Finder are exported from the centralised REDCap database where BCNTB data are stored, reflecting the most up-to-date clinical and sample data available at the time. Using Sample Finder, researchers can instantly see how many cases match their defined cohort and view a summary report in the form of a graphical breakdown of the

matching cases' donor and tumour statistics. In case of successful biospecimen and data request applications, researchers are invited to return the data generated as part of their study to the Bank after the study results enter the public domain. The Biospecimen and Data Request System is available at <https://breastcancer.org/breast-cancer-research/breast-cancer-now-tissue-bank/how-apply-tissue-bank> accessed on 15 February 2023.

Bioinformatics Portal. The Bank hosts its own bioinformatics portal, implemented using open-source Python and R packages [21], that includes Analytics Hub, a data analytics platform. The portal acts as a central hub, giving researchers access to a list of associated publications, user guides and the Bank's contact information.

Analytics Hub, powered by a custom version of SNPnexus [22], provides an environment on which the output of Whole Genome Sequencing pipelines (e.g., FASTQ and VCF files), clinical and omics data returned by studies that used BCNTB-BCI biospecimens (e.g., the Spatial Characterisation [23] and the 100K GP cohort datasets) and publicly available data from large-scale studies (e.g., the Cancer Genome Atlas (TCGA) [24], the Cancer Cell Line Encyclopedia (CCLE) [25] and the International Cancer Genome Consortium (ICGC) [26] datasets) can be viewed, filtered, analysed and incorporated into research projects and grant applications. Analytics Hub has bidirectional links to Sample Finder, i.e., upon entering a query in Analytics Hub, the platform also checks the availability of biospecimens and data from BCNTB-BCI donors with the same characteristics on Sample Finder, for which researchers can submit an EOI to the bank; inversely, upon entering a query in Sample Finder, the tool also checks the availability of data for patients from other studies with the same characteristics of interest, which researchers can explore on Analytics Hub. The Bioinformatics Portal is available at <https://bcntb.bcc.qmul.ac.uk/home> accessed on 15 February 2023.

3. Results

3.1. BCNTB-BCI Infrastructure

Figure 1 shows the overview of the described BCNTB-BCI infrastructure.

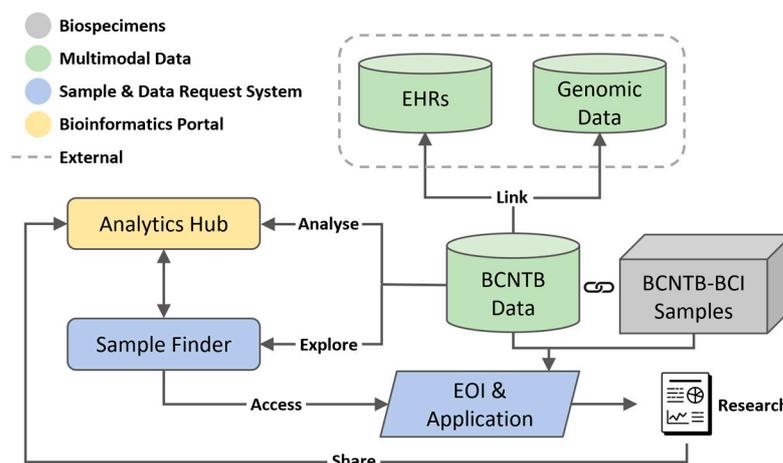


Figure 1. Overview of BCNTB-BCI infrastructure showing the centralised REDCap database where BCNTB data are stored, linkage to EHRs from BH and genomic data from 100K GP, the integrated Biospecimen and Data Request System and Bioinformatics Portal.

3.2. BCNTB-BCI Cohort Overview

To date, BCNTB-BCI has collected over 42,000 biospecimens and data from 2400 breast cancer patients and individuals without the disease (6% of donors approached for consent have dissented). Longitudinal liquid samples and derivatives, tissue samples and cell cultures are available for 17%, 99% and 21% of BCNTB-BCI donors, respectively, with

samples collected at the time of diagnosis and through the treatment cycles until the last follow-up or death.

Linkage of BCNTB-BCI data to EHRs allows a quick inspection of the breast cancer landscape in our consented cohort. We selected a subset of BCNTB-BCI donors (data extracted on 22 September 2022) who have been followed up for at least 5 years or until death (whichever is earlier) and for whom age at diagnosis, ethnicity, disease and survival status data are available and independently verified throughout this period ($n = 1066$). Figures 2–4 show the evolution of disease status in this subset of donors over 5 years of follow-up (the period during which risk of breast cancer recurrence is high), the prevalence and co-occurrence of conditions that are often comorbid with breast cancer in these patients, and their age at diagnosis of primary breast cancer and ethnicity, respectively. Our results suggest that breast cancer screening uptake in the consented population of East London covered by BH is low, resulting in many cases presenting at later stages of the disease. For instance, the de novo metastatic breast cancer rate is at least 2% and the incidence of recurrence is the highest in the first year after diagnosis (Figure 2). Oestrogen receptor (ER) negative breast cancer cases make up the majority of recurrence cases diagnosed in this period. Additionally, the results show that the most common comorbid condition is cardiovascular disease with a prevalence rate of 38%, followed by hypertension and rheumatic disease (Figure 3). This is in agreement with statistics reported in numerous other studies that investigated the prevalence of comorbidities in breast cancer patients (e.g., see [27]). However, our results also highlight the challenges in breast cancer detection and treatment in the multi-ethnic population of East London, where over 20% and over 30% of patients are Black and Asian [14], respectively. Overall survival of patients, comprised of relapse-free (the ‘Primary’ category) and secondary breast cancer (the ‘Recurrence’ category) cases, is worse in these ethnic groups. For example, the 5-year overall survival rate for our consented Black group is 78%, which is lower than the rate for all BCNTB-BCI donors (84%, Figure 2) and the national average of 85% [1]. Similarly, the age at diagnosis is younger particularly in the Asian and Black groups (Figure 4). In addition, the results highlight some of the challenges in breast cancer management in this region. For example, the comorbidity burden is high, with 14% of patients experiencing three or more comorbidities (Figure 3).

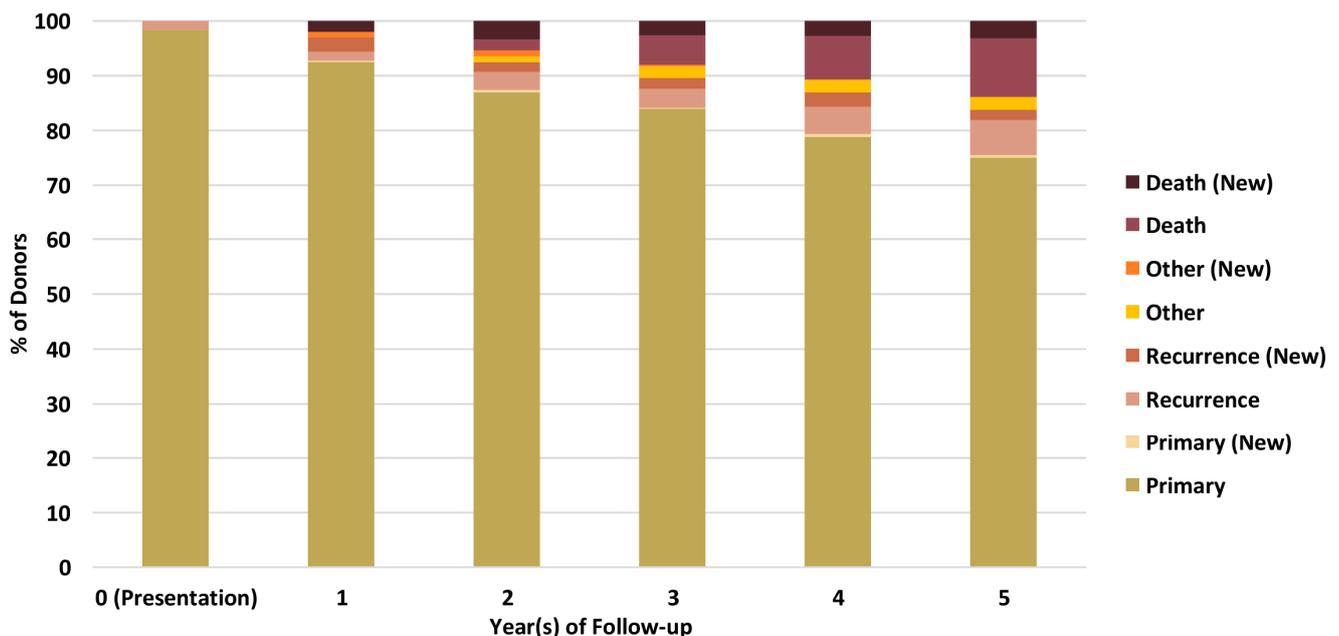


Figure 2. Evolution of disease status in BCNTB-BCI donors over 5 years of follow-up (the ‘Other’ category consists of cosmetic and high-risk cases; new and existing cases at the end of each follow-up year are shown separately).

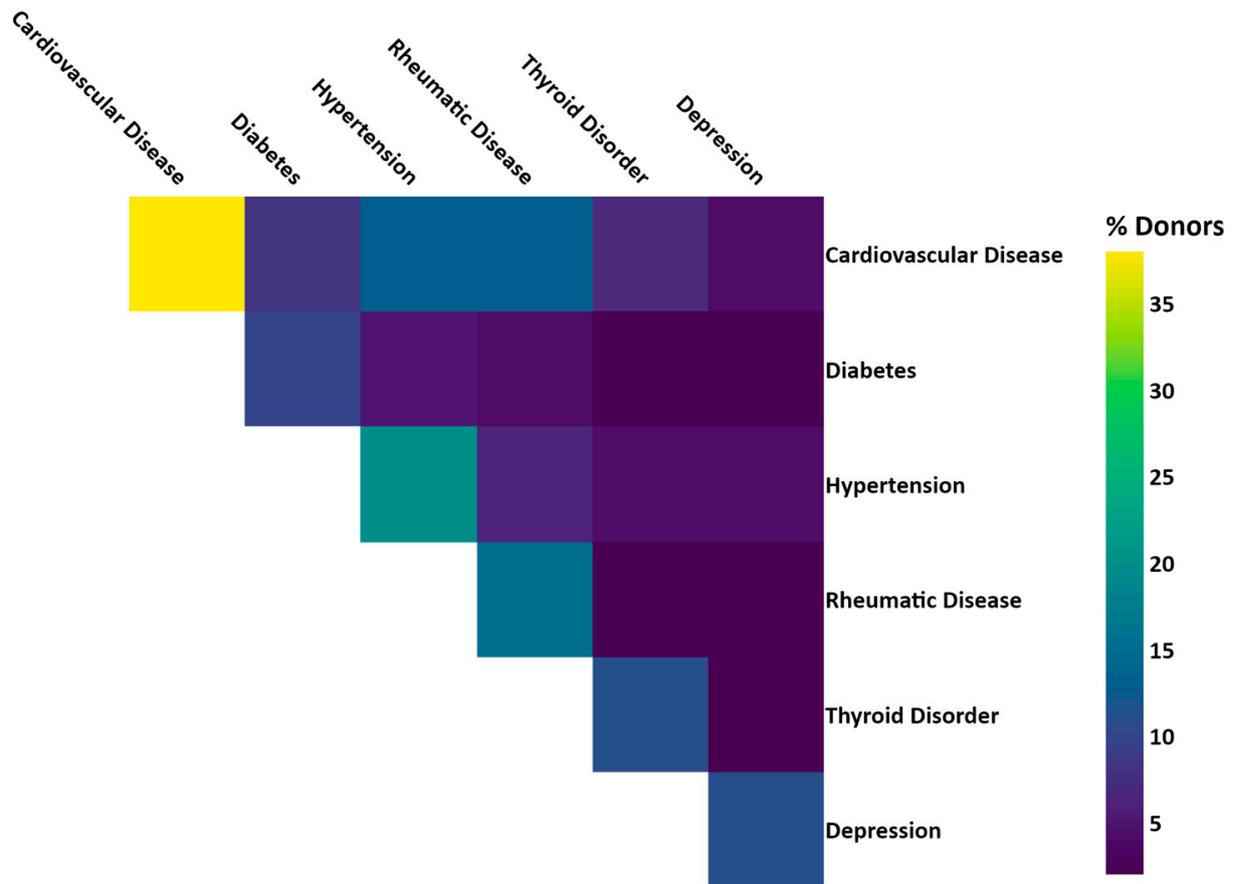


Figure 3. Prevalence and co-occurrence of conditions comorbid with breast cancer in BCNTB-BCI patients (the numbers are calculated based on the ICD-10 codes used to compute the Elixhauser comorbidity index [28]).

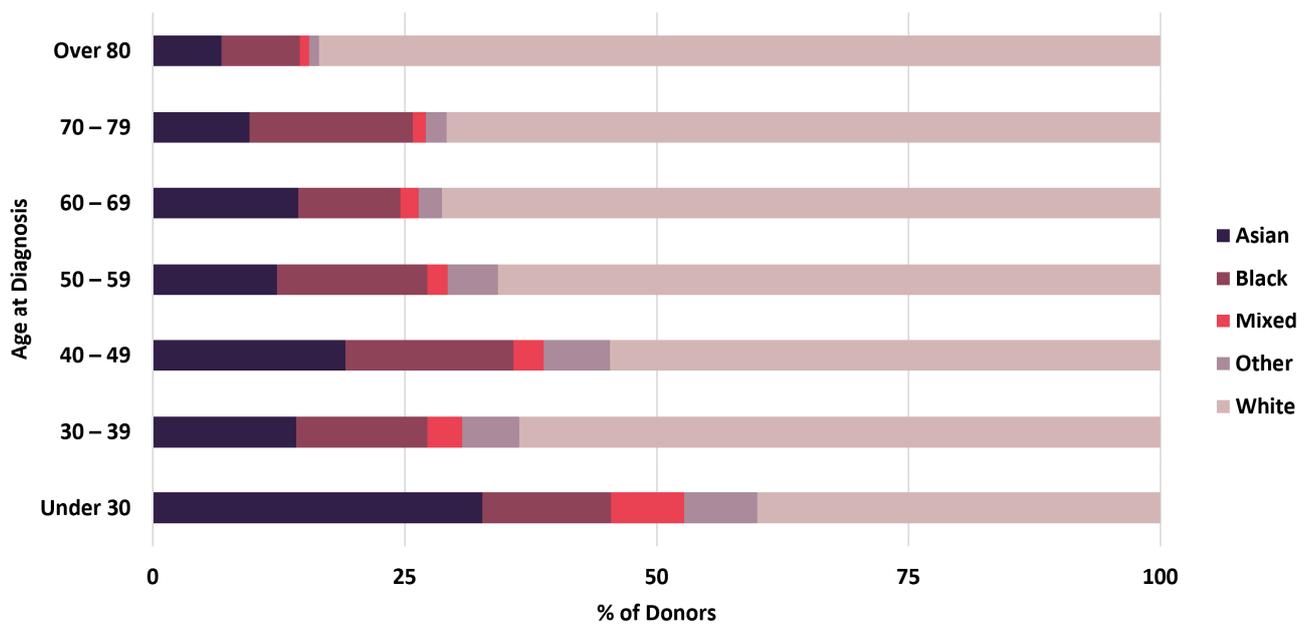


Figure 4. Age at diagnosis of primary breast cancer and ethnicity of BCNTB-BCI patients.

3.3. Clinical and Molecular Patient Journey Narratives

The infrastructure of our cancer biobank facilitates the generation of clinical and molecular patient journey narratives, which are powerful tools for precision medicine. Figure 5 shows the clinical patient journey narrative for a breast cancer patient (referred to as Patient A) over 5 years (denoted by I to V in the figures) based on their biobank data from BCNTB-BCI. These data are mostly structured, which eases visualisation and interpretation of patient journey narratives.

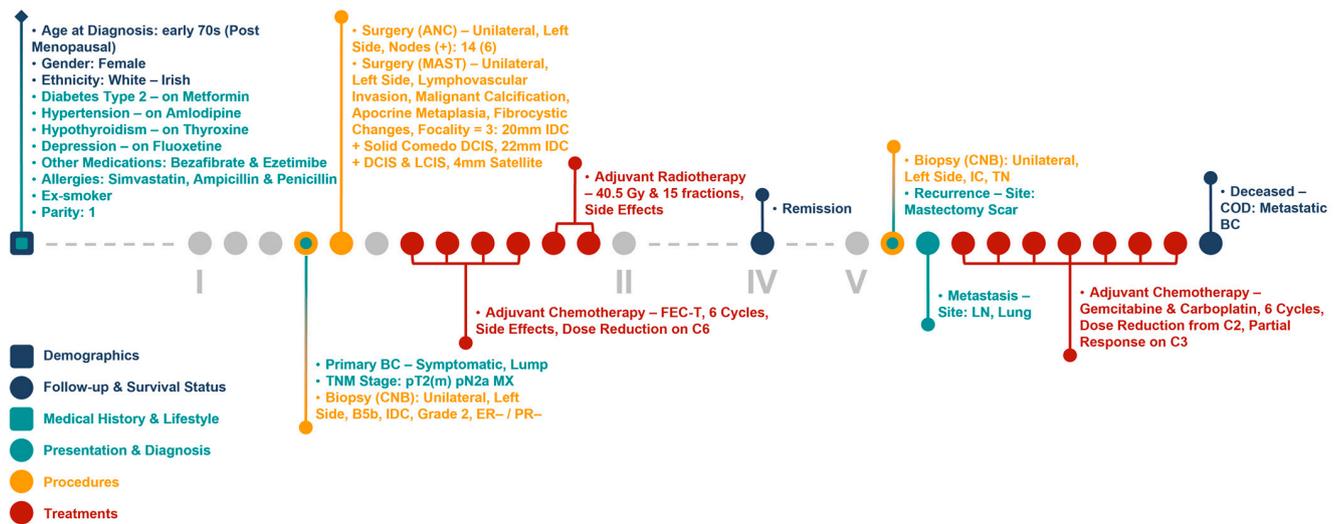


Figure 5. Example patient journey based on BCNTB-BCI data.

Figures 6 and 7 show the clinical patient journey narratives for Patient A based on the structured and unstructured parts (specifically imaging reports) of their EHRs from BH, respectively. EHRs differ from biobank data in terms of what information is recorded, and when and how this information is recorded: compared to BCNTB-BCI data, EHRs from BH provide a better coverage of clinical events, which makes them suitable for certain studies (e.g., comorbidity analysis); however, they are not necessarily cancer-specific, and therefore, require filtering before they can be used to visualise cancer patient trajectories. In BCNTB-BCI data, clinical events appear sooner and with a single date, whereas in EHRs from BH, codes may have been entered multiple times with different dates. As a cancer-oriented biobank, BCNTB records cancer-related information explicitly and mostly in a structured format, while in EHRs, the same information may not be directly available or are recorded in an unstructured format. For instance, BCNTB-BCI records age at diagnosis, family history of cancer and treatment regimen, including the response and any side effects, explicitly. In contrast, using EHRs from BH, we calculated age at diagnosis based on date of birth and date of the earliest instance of the relevant diagnosis code. Likewise, we filtered family history data by relation (first-degree relative or otherwise) and disease to obtain cancer-specific family history information. Similarly, drugs that are part of the same regimen are not grouped together, and the response to treatment and any side effects are usually mentioned in free-text reports; we extracted and aggregated this information.

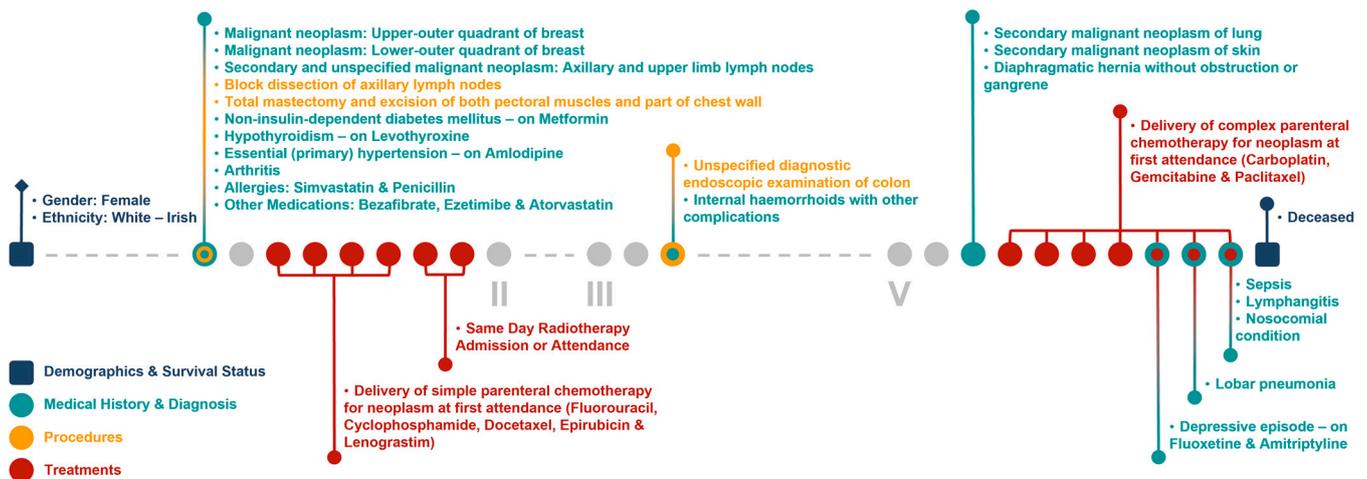


Figure 6. Example patient journey based on the structured parts of EHRs from BH (non-cancer-related information are omitted; ICD-10 and OPCS-4 descriptions are shown for diagnoses and procedures, respectively; drugs are shown next to the condition they were prescribed for).

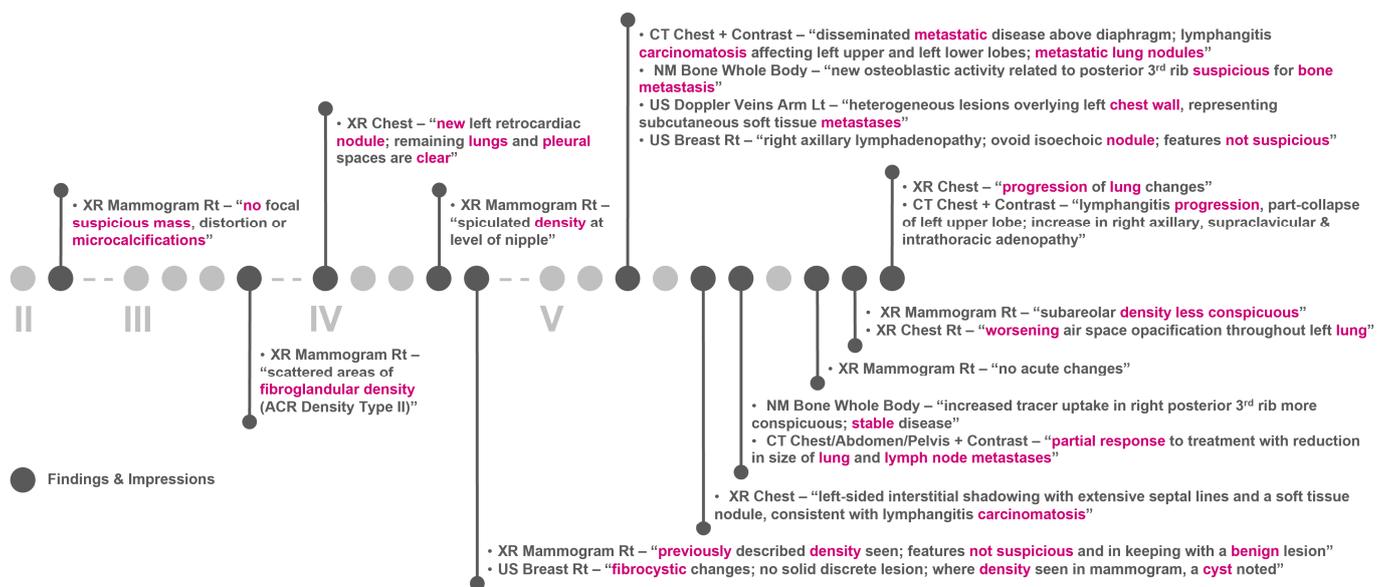
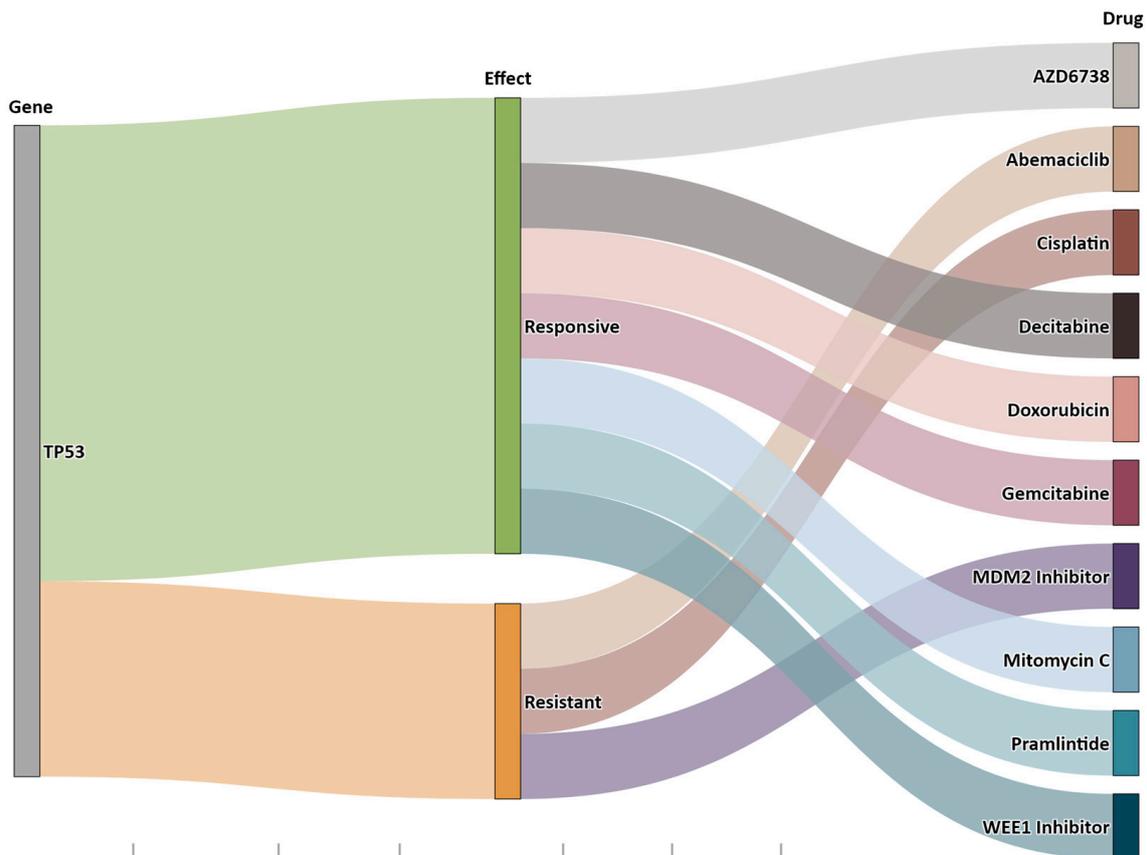


Figure 7. Example patient journey based on imaging reports in EHRs from BH (non-cancer-related information are omitted; keywords highlighted by the NLP algorithm are shown in purple; reports are paraphrased and summarised).

Using 100K GP data, we identified somatic variants for Patient A using the Analytics Hub. These variants were used as input query set to the Cancer Genome Interpreter (CGI, release 28) [29], which produced a list of potential therapeutically actionable candidates. In the case of this patient, TP53 D281N variant was listed as a biomarker of drug response. Figure 8 shows the drug sensitivity profiles for this variant. According to EHRs from BH, the patient was administered a combination chemotherapy regimen of carboplatin, gemcitabine and paclitaxel, and according to BCNTB-BCI data, the patient had an initial partial response followed by disease progression and died shortly after. The predictive power of TP53 mutations has not yet been sufficiently validated to permit conclusions on the pharmacogenomics to be drawn. Genomic insights, in conjunction with information from other sources (e.g., biobank data and EHRs), not only provide researchers with richer datasets, but also guide clinicians in their diagnosis and management decision-making process.



Alteration	Biomarker	Drug	Disease	Association	Evidence	Source
TP53 (D281N)	TP53 Oncogenic Mutation	AZD6738	B Cell Lymphoma	Responsive	Early trials	NCT01955668
TP53 (D281N)	TP53 Oncogenic Mutation	Abemaciclib	Breast Adenocarcinoma	Resistant	Early trials	PMID:27217383
TP53 (D281N)	TP53 Oncogenic Mutation	Cisplatin	Female / Male Germ Cell Tumour	Resistant	Early trials	PMID:27646943
TP53 (D281N)	TP53 Oncogenic Mutation	Decitabine	Acute Myeloid Leukemia	Responsive	Early trials	PMID:27959731
TP53 (D281N)	TP53 Oncogenic Mutation	Doxorubicin	Bladder	Responsive	Pre-clinical	PMID:27397505
TP53 (D281N)	TP53 Oncogenic Mutation	Gemcitabine	Bladder	Responsive	Pre-clinical	PMID:27397505
TP53 (D281N)	TP53 Oncogenic Mutation	MDM2 Inhibitor	Liposarcoma	Resistant	Early trials	PMID:23084521; ASCO 2015 Abstract:10564
TP53 (D281N)	TP53 Oncogenic Mutation	Mitomycin C	Bladder	Responsive	Pre-clinical	PMID:27397505
TP53 (D281N)	TP53 Oncogenic Mutation	Pramlintide	Thymic	Responsive	Pre-clinical	PMID:25409149
TP53 (D281N)	TP53 Oncogenic Mutation	WEE1 Inhibitor	Head and Neck	Responsive	Pre-clinical	PMID:25125259
TP53 (D281N)	TP53 Oncogenic Mutation	WEE1 Inhibitor	Ovary	Responsive	Early trials	PMID:27998224

Figure 8. Example of a patient’s sequencing data from 100K GP analysed on Analytics Hub (only gene-drug pairs where alteration match is complete are shown).

4. Discussion

We adopted a dynamic infrastructure using REDCap for the management of BCNTB data and provision of its biospecimens. We implemented and customised measures that improve the quality (correctness, completeness, consistency and currency or timeliness) of

the Bank's data. Additionally, mapping our data to a standard biomedical ontology (the SNOMED-CT concepts) facilitated their integration and interoperability with EHRs and genomic data. This resulted in a rich longitudinal multimodal dataset that captures the breast cancer landscape in a multi-ethnic population and highlights the shortcomings in detection, management and treatment of breast cancer in the local area. Previous studies examining breast cancer presentation in this geographical region carried out surveys (via patient questionnaires and interviews) to collect data, which rely on large samples that are representative of the target population and high response rates [30]. BCNTB-BCI data can supplement, validate and guide such studies.

The BCNTB integrated Biospecimen and Data Request System offers real-time information on the availability of biospecimens and data to researchers, while simplifying the data sharing process to maximise data reuse. The BCNTB integrated Bioinformatics Portal enables the research community not only to access high-quality annotated biospecimens, but also to interrogate the associated data. Analysing various types of data available for each patient is key to implementing personalised diagnostic, prognostic and therapeutic strategies. It reduces research time and complexity, while the seamless data exchange and navigation between Sample Finder and Analytics Hub helps overcome the barriers due to lack of data compatibility. This created a domain-specific repository of breast cancer data and tools that can be used to generate clinical and molecular patient journey narratives. Such narratives offer context and a cross-functional view of patient experience that cannot be achieved without employing these tools and by interrogating data from either source alone, or engage all stakeholders to identify gaps in breast cancer care.

Our observations are limited to a consented subset of donors followed up for over 5 years, and therefore, need to be validated in a larger unbiased cohort. Another limitation is the impact of donors who are lost to follow-up. Linkage of BCNTB-BCI data to primary care EHRs from the Discovery East London programme [31] is currently in progress. This will enable us to generate near-complete data for more comprehensive clinical patient journey narratives by enriching the demographics, lifestyle, comorbidity and mortality data that have so far been recorded in hospital visits.

5. Conclusions

It is important to note some of the challenges and lessons learned from our work. The underlying infrastructure of a biobank must offer scalability and interoperability, allowing data growth and flow. Biobanking staff at multiple sites need a user-friendly system that supports entry, querying, management and auditing of clinical and sample data. Researchers need a system that enables them to search and request biospecimens and data. They also need the tools by which data generated from the bank's biospecimens as well as related initiatives can be analysed. Importantly, the infrastructure of a biobank should evolve in response to feedback from clinicians and researchers.

A challenge faced by multisite biobanks is the variation in terminology used both within and between collaborating centres. This is an issue, especially for retrospective collections, that must be resolved through the implementation of a controlled vocabulary for all the biobank's operating centres using international standard biomedical ontologies. Another challenge is managing the collections from legacy sites and facilitating the onboarding of new sites. Furthermore, extracting information from EHRs requires domain experts to be involved from the start of any biobanking project. At BCI, we work closely with our NHS Trust to develop knowledge and best practices for transforming health records into breast cancer research-ready data. This, however, poses a challenge when patients are lost to follow-up at BH.

Data generated from a biobank's biospecimens are its legacy. Sharing these data helps avoid duplication of effort and financial resources for researchers and depletion of rare biospecimens, and allows researchers to conduct preliminary analysis and validation of ongoing projects. Data sharing is vital for translation of research to patient benefit. Biobanks can be a powerful catalyst in tackling resistance to data sharing.

We present BCNTB-BCI as a dynamic cancer biobank with an infrastructure that includes several features to help tackle these challenges in breast cancer research. They include: Implementation of a flexible yet robust data management system; linkage to other health data sources; and integrated data access and analytics tools. We demonstrated how these features improve data quality, increase interoperability, encourage data reuse and accelerate research. Specifically, we showcased an example where such a biobanking ecosystem can aid in predicting the health trajectory of breast cancer patients and identifying potential actionable opportunities for intervention.

Author Contributions: Conceptualisation, M.A. (Maryam Abdollahyan), E.G. and C.C.; methodology, M.A. (Maryam Abdollahyan), E.G. and C.C.; software, M.A. (Maryam Abdollahyan), E.G., M.A. (Millahat Asif) and J.O.; validation, R.B.-M.; formal analysis, M.A. (Maryam Abdollahyan) and E.G.; investigation, M.A. (Maryam Abdollahyan) and E.G.; resources, L.J.J. and C.C.; data curation, M.A. (Maryam Abdollahyan), E.G.; writing—original draft preparation, M.A. (Maryam Abdollahyan) and E.G.; writing—review and editing, M.A. (Maryam Abdollahyan), E.G., M.A. (Millahat Asif), J.O., R.B.-M., S.J., C.C. and L.J.J.; visualisation, M.A. (Maryam Abdollahyan), E.G. and M.A. (Millahat Asif); supervision, C.C.; project administration, C.C.; funding acquisition, L.J.J. and C.C. All authors have read and agreed to the published version of the manuscript.

Funding: BCNTB is funded by the Breast Cancer Now charity (grant number TB2022BAR). J.O. is supported by the Barts Charity (grant number MGU0344).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the NHS Research Ethics Committee (reference number 21/EE/0072).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. Patient data are not publicly available due to ethical restrictions.

Acknowledgments: This paper is dedicated to all the individuals who have donated to BCNTB. We are grateful to members of BCNTB-BCI (Rachel Nelan, Jennifer McGuinness, Jenny Gomm and Iain Goulding) and other BCNTB centres (led by Angela Cox and Valerie Speirs) for their help in setting up the framework for collection and distribution of biospecimens.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cancer Research UK. Breast Cancer Statistics. Available online: <https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer> (accessed on 1 October 2022).
2. Pan, H.; Gray, R.; Braybrooke, J.; Davies, C.; Taylor, C.; McGale, P.; Peto, R.; Pritchard, K.I.; Bergh, J.; Dowsett, M.; et al. 20-year risks of breast-cancer recurrence after stopping endocrine therapy at 5 years. *N. Engl. J. Med.* **2017**, *377*, 1836–1846. [CrossRef] [PubMed]
3. Breast Cancer Now. Almost One Million Women in UK Miss Vital Breast Screening Due to COVID-19. Available online: <https://breastcancernow.org/about-us/media/press-releases/almost-one-million-women-in-uk-miss-vital-breast-screening-due-covid-19> (accessed on 1 October 2022).
4. World Health Organization. *The International Statistical Classification of Diseases and Health Related Problems, 10th Revision, Volume 1, Tabular List*, 5th ed.; WHO: Geneva, Switzerland, 2016.
5. NHS Digital. OPCS Classification of Interventions and Procedures. Available online: https://datadictionary.nhs.uk/supporting_information/opcs_classification_of_interventions_and_procedures.html (accessed on 1 October 2022).
6. National Institutes of Health. Systematized Nomenclature of Medicine—Clinical Terms. Available online: <https://www.nlm.nih.gov/healthit/snomedct> (accessed on 1 October 2022).
7. Breast Cancer Now. Breast Cancer Now Tissue Bank. Available online: <https://breastcancernow.org/breast-cancer-research/breast-cancer-now-tissue-bank> (accessed on 1 October 2022).
8. Felmeister, A.S.; Rivera, T.J.; Masino, A.J.; Resnick, A.C.; Pennington, J.W. Scalable biobanking: A modular electronic honest broker and biorepository for integrated clinical, specimen and genomic research. In Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Washington, DC, USA, 9–12 November 2015; pp. 484–490.
9. Kinkorová, J. Biobanks in the era of personalized medicine: Objectives, challenges, and innovation. *EPMA J.* **2016**, *7*, 1–12. [CrossRef] [PubMed]

10. Olson, J.E.; Bielinski, S.J.; Ryu, E.; Winkler, E.M.; Takahashi, P.Y.; Pathak, J.; Cerhan, J.R. Biobanks and personalized medicine. *Clin. Genet.* **2014**, *86*, 50–55. [CrossRef] [PubMed]
11. NHS North East London Clinical Commissioning Group. East London Patient Record. Available online: <https://northeastlondonccg.nhs.uk/about-us/elpr.htm> (accessed on 1 July 2021).
12. Harris, P.A.; Taylor, R.; Thielke, R.; Payne, J.; Gonzalez, N.; Conde, J.G. Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. *J. Biomed. Inform.* **2009**, *42*, 377–381. [CrossRef] [PubMed]
13. Harris, P.A.; Taylor, R.; Minor, B.L.; Elliott, V.; Fernandez, M.; O’Neal, L.; McLeod, L.; Delacqua, G.; Delacqua, F.; Kirby, J.; et al. The REDCap consortium: Building an international community of software platform partners. *J. Biomed. Inform.* **2019**, *95*, 103208. [CrossRef] [PubMed]
14. Barts Health NHS Trust. Fast facts. Available online: <https://www.bartshealth.nhs.uk/fast-facts> (accessed on 1 October 2022).
15. Somerset NHS Foundation Trust. Somerset Cancer Register. Available online: <https://www.somersetft.nhs.uk/somerset-cancer-register> (accessed on 1 October 2022).
16. NHS Digital. Data Collections and Data Sets. Available online: <https://digital.nhs.uk/data-and-information/data-collections-and-data-sets> (accessed on 1 October 2022).
17. Centre for Genomic Regulation and European Bioinformatics Institute. The European Genome-Phenome Archive. Available online: <https://ega-archive.org> (accessed on 1 October 2022).
18. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **2015**, *526*, 68. [CrossRef] [PubMed]
19. National Cancer Intelligence Network. The National Cancer Registration and Analysis Service. Available online: <http://www.ncin.org.uk> (accessed on 1 October 2022).
20. Grinberg, M. *Flask Web Development: Developing Web Applications with Python*; O’Reilly Media, Inc.: Sebastopol, CA, USA, 2018.
21. Gentleman, R.C.; Carey, V.J.; Bates, D.M.; Bolstad, B.; Dettling, M.; Dudoit, S.; Ellis, B.; Gautier, L.; Ge, Y.; Gentry, J.; et al. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol.* **2004**, *5*, 1–16. [CrossRef] [PubMed]
22. Oscanoa, J.; Sivapalan, L.; Gadaleta, E.; Dayem Ullah, A.Z.; Lemoine, N.R.; Chelala, C. SNPnexus: A web server for functional annotation of human genome sequence variation (2020 update). *Nucleic Acids Res.* **2020**, *48*, W185–W192. [CrossRef] [PubMed]
23. Gadaleta, E.; Fourgoux, P.; Pirr6, S.; Thorn, G.J.; Nelan, R.; Ironside, A.; Rajeeve, V.; Cutillas, P.R.; Lobley, A.E.; Wang, J.; et al. Characterization of four subtypes in morphologically normal tissue excised proximal and distal to breast cancer. *NPJ Breast Cancer* **2020**, *6*, 1–12. [CrossRef] [PubMed]
24. National Cancer Institute. The Cancer Genome Atlas. Available online: <https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga> (accessed on 1 October 2022).
25. Broad Institute of Harvard and MIT. The Cancer Cell Line Encyclopedia. Available online: <https://sites.broadinstitute.org/ccl> (accessed on 1 October 2022).
26. International Cancer Genome Consortium. ICGC Data Portal. Available online: <https://dcc.icgc.org> (accessed on 10 February 2023).
27. Ng, H.S.; Vitry, A.; Koczwara, B.; Roder, D.; McBride, M.L. Patterns of comorbidities in women with breast cancer: A Canadian population-based study. *Cancer Causes Control* **2019**, *30*, 931–941. [CrossRef] [PubMed]
28. Elixhauser, A.; Steiner, C.; Harris, D.R.; Coffey, R.M. Comorbidity measures for use with administrative data. *Med. Care* **1998**, *36*, 8–27. [CrossRef] [PubMed]
29. BBGLab. The Cancer Genome Interpreter. Available online: <https://www.cancergenomeinterpreter.org> (accessed on 1 October 2022).
30. Forbes, L.J.L.; Atkins, L.; Thurnham, A.; Layburn, J.; Haste, F.; Ramirez, A.J. Breast cancer awareness and barriers to symptomatic presentation among women from different ethnic groups in East London. *Br. J. Cancer* **2011**, *105*, 1474–1479. [CrossRef] [PubMed]
31. Discovery East London. Discovery Data Service. Available online: <https://www.discoverydataservice.org> (accessed on 10 February 2023).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.